Human-Like Moral Decisions by Reinforcement Learning Agents

Ali Shiravand (ali.shiravand@ens.fr)
Department of Cognitive Studies, École Normale Supérieure-PSL University,
29 Rue d'Ulm, 75005 Paris, France

Jean-Baptiste André (jean-baptiste.andre@ens.fr)
Institut Jean Nicod, Department of Cognitive Studies, École Normale Supérieure-PSL University,
29 Rue d'Ulm, 75005 Paris, France

Abstract

Human moral judgments are both precise, with clear intuitions about right and wrong, and at the same time obscure, as they seem to result from principles whose logic often escapes us. The development of Artificial Intelligence (AI) applications requires an understanding of this subtle logic if we are to embed moral considerations in artificial systems. Reinforcement Learning (RL) algorithms have emerged as a valuable interactive tool for investigating moral behavior. However, being value-based algorithms, they face difficulty when it comes to explaining deontological, non-consequentialist moral judgments. Here, in a multi-agent learning scenario based on the Producer-Scrounger Game, we show that RL agents can converge towards *apparently* non-consequentialist outcomes, provided the algorithm accounts for the temporal value of actions. The implications of our findings extend to integrating morality into AI agents by elucidating the interplay between learning strategies, characteristics for accounting temporal values, and methods of considering the opponent's payoff.

Keywords: Moral Behavior; Multi-Agent Reinforcement Learning; Consequentialism; Human-AI Interaction; Behavioral Game Theory

Introduction

Human judgments of right and wrong exhibit nuanced variation. Although we support equality in resource allocation, we generally tolerate significant income disparities between different professions (Starmans et al., 2017). We uphold the inviolability of private property while condemning an art lover's act of destroying a newly acquired painting. We uphold the importance of obeying laws, yet refuse to obey them when they seem unjust. Our moral judgments are precise, with clear intuitions about right and wrong in particular scenarios. Yet they also remain enigmatic, seemingly derived from disparate principles whose underlying logic often eludes our grasp (Haidt, 2001; Hauser et al., 2007). This diversity manifests itself in some instances as an orientation toward taking into account the immediate consequences of one's actions, i.e., being consequentialist, while in others embracing apparently deontological perspectives that eschew cost-benefit analysis in favor of adherence to unwavering ethical principles (Haidt, 2001; Hauser et al., 2007).

Numerous moral theories, echoing Kant's categorical imperative, posit that human moral judgments operate akin to contractual agreements (Levine et al., 2020; Gauthier, 1987; André et al., 2022). The principle of the golden rule, encapsulating the essence of reciprocity—"treat others as you would like to be treated"—serves as a cornerstone. In a workplace setting, for instance, the golden rule might manifest in the form of fair and respectful treatment of colleagues. If each individual universally adopts this principle, the work environment is likely to become more cooperative, supportive, and conducive to productivity. This example illustrates how moral rules, when universalized, can serve as foundational guides for ethical behavior across various situations and societal contexts.

Even with the golden rule as a guiding principle, it doesn't imply that moral judgments are simple or consistent. Rather they are subtle and often highly context-dependent. For instance, imagine a scenario where a friend confides in you about a personal struggle they are facing but explicitly requests that you do not share this information with anyone else. In this context, a moral judgment is required concerning the value of honesty. While the principle of honesty is generally upheld, the specific context creates a moral dilemma. Revealing your friend's confidential information could breach trust and potentially harm your friend, suggesting a nuanced consideration of other ethical principles, such as loyalty and the avoidance of harm. This shows that moral judgments are not solely determined by a fixed set of principles but are influenced by the context in which they arise. The balancing act between honesty and other ethical considerations highlights the dynamic and situation-dependent nature of human moral reasoning.

Considering all the complex nature of human moral judgments, nowadays, Artificial Intelligence (AI) is becoming more and more entangled in human's everyday life. We need to keep in mind that the development of AI systems should be done with consideration of ethical, legal, and social implications. All these aspects raise the "AI alignment problem," which refers to the challenge of ensuring that advanced artificial intelligence (AI) systems act in accordance with human values, goals, and intentions (Yudkowsky, 2016). As AI systems become more sophisticated, there is a growing concern that they might not align with human values or that their objectives may deviate from what is desired by their human creators. The problem of AI alignment becomes particularly critical when considering systems that operate autonomously or in complex and uncertain environments. If the goals and values of AI systems are not properly aligned with those of humans, there is a risk of unintended consequences

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

or even potentially harmful outcomes. The challenge of AI alignment has both technical and philosophical dimensions. From a technical perspective, it requires designing AI systems that accurately understand and prioritize human values, adapt to changes in those values, and avoid undesirable behavior. From a philosophical perspective, it involves addressing questions about how to define and represent human values, as well as ethical considerations related to the use of AI. (Christian, 2021; Sorensen et al., 2023).

Reinforcement learning (RL) algorithms have been widely used to model human value-based decisions (Schultz et al., 1997; O'Doherty et al., 2004; Pessiglione et al., 2006). Through interacting with the environment and receiving external feedback in the form of rewards or punishments, RL agents learn a policy to decide (Sutton & Barto, 2018). Alongside their applications in modeling individual decisions, these types of interactive algorithms have also been used in game theory to study human social and moral behavior (Horita et al., 2017; Nguyen et al., 2020; Herlau, 2022). In the same direction, Ecoffet and Lehman (2021) discussed the training of ethical agents through reinforcement by rewarding correct behavior under certain moral theories, suggesting the potential for RL to address moral uncertainty (when we are uncertain about the outcome of our moral decisions).

The aforementioned studies assume an objective view of the game by the agent, meaning that the value of the outcome of the game is perceived the same across agents. However, a subjective representation of the game for each agent might drastically change the outcome and there are several studies that argue the existence of subjective utility functions in humans (e.g. Frey and Stutzer (2002); Shadmehr et al. (2019)). Conitzer et al. (2017) discussed game-theoretic representation schemes as one of the most prominent representations of moral dilemmas. Nevertheless, they focused on moral solution representation in their study, not a subjective view of the game. Here in this study, we argue that the subjective representation of the game (e.g. deontological vs. utilitarian view over the game by each agent) can significantly affect the outcome in a game-theoretic setup. We claim that even in a complete value-based game, the perspective of the learning agents will determine the outcome.

Building upon earlier research investigations on model-free and model-based processes in consequentialist and deontological moral judgments (Cushman, 2013; Crockett, 2013), this study aims to present the implementation of an agreement-based principle within model-free RL algorithms (Sutton & Barto, 2018) and to investigate its potential to yield moral decisions similar to those observed in human contexts by using a simple strategic game. Focusing on the evaluation of moral judgments regarding the rights to use material resources, a domain where distinct moral principles appear to guide human behavior, we delve into the complexities of individual decision-making.

In our research, individuals navigate two distinct roles: (i) where they can actively produce resources, and (ii) where they can choose to take resources produced by others, knowing that the benefits to them exceed the costs to others. The study includes two modes of learning, a "selfish" mode and a "utilitarian" mode. The analysis reveals three potential outcomes depending on individuals' learning modes: (i) an unproductive selfish outcome, characterized by resource appropriation without production effort; (ii) a utilitarian outcome, in which individuals engage in both resource production and consumption; and (iii) a seemingly deontological outcome, in which individuals produce resources but refrain from taking them, even when doing so could be both personally and socially beneficial. We argue that these outcomes reflect the diverse range of moral judgments inherent in human perceptions of material resource ownership.

Our simulations extend beyond individual decision-making to capture the broader spectrum of social trust prevalent in human societies. In scenarios where trust is low, individuals prioritize personal gain in all situations, culminating in a society characterized by power relations. Intermediate levels of trust result in selective cooperation; individuals refrain from taking resources from others, but fall short in collective resource production, leading to a societal archetype that resembles a "bourgeois" private property framework. Conversely, high levels of trust foster universal cooperation in all situations, even encouraging resource production in the absence of personal benefit. This cooperative ethos culminates in a society characterized by solidarity and sharing, embodying a paradigm where contributions match abilities and benefits match needs. Through this study, we provide insights into the dynamic interplay between moral decision-making, learning modes, and societal trust in the domain of interactive AI algorithms.

Methods

Game Design

To simulate different social scenarios in our study, we used the Producer-Scrounger Game (Figure 1), which retains the basic structure of the original two-player paradigm. During the game, the Producer chooses between producing (P) and not producing (NP) a resource, and the Scrounger chooses between Taking (T) and Leaving (L) the resource. If the Producer chooses not to produce (NP), the Scrounger's decision does not affect the outcome, resulting in both players receiving equally low rewards. We call this situation the "Unproductive" Outcome. This situation corresponds to the Subgame Perfect Nash Equilibrium (SPNE), a state expected by rational agents in the non-repeating context of the game (Selten & Bielefeld, 1988).

However, if the Producer chooses to produce (P) and the Scrounger leaves (L), the Producer receives the intended reward for their effort, and the Scrounger's reward remains unchanged from the NP scenario. In this context, the Producer's effort does not result in any benefit to the Scrounger, which represents the "Private Property" Outcome.

If the Producer chooses to produce (P) and the Scrounger

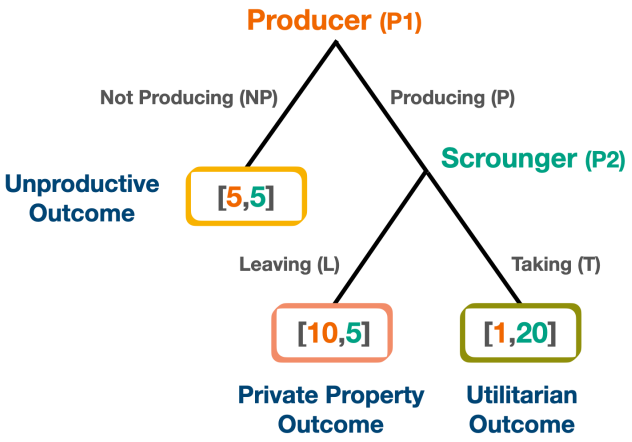

Figure 1: Extensive form of the Producer-Scrounger Game

chooses to take (T), the Producer receives a lower payoff than in the Uproductive Outcome, while the Scrounger receives a higher payoff than what the Producer receives in the Private Property Outcome. This result characterizes the "Utilitarian" Outcome, a state in which the sum of payoffs in the Producer-Scrounger Game is maximized.

An important difference between the Producer-Scrounger Game and the Trust Game (Berg et al., 1995), is the P-L condition. In the standard version, the Scrounger receives more than the Unproductive Outcome. In our adapted Producer-Scrounger Game, however, the Scrounger's payoff in this condition remains equal to the Unproductive Outcome. These nuanced differences serve as critical components in our exploration of social interactions and decision dynamics in the context of our research.

Agent Design

In this study, we used a multi-agent simulation approach to investigate the dynamic patterns and converged states of social behavior. To model the adaptive nature of human social interactions, we used model-free RL agents in the context of the Producer-Scrounger Game.

In our simulation environment, there are two agents, each undergoing an autonomous learning process. During each trial, the agents are randomly assigned to the roles in the game (the Producer or the Scrounger), indicating their states in the game. Based on their role, each player has two possible actions: 1) for the Producer, the choice is between P or NP, and 2) for the Scrounger, the choices are T or L. As shown in Equation 1, each agent learns the values associated with these four possible actions in the game using the Rescorla-Wagner Rule (Rescorla, 1972).

$$Q_t(s_t,a_t) = (1-\alpha).Q_{t-1}(s_t,a_t) + \alpha.r_t \tag{1}$$

In Equation 1, s_t represents the state (role) of the agent in the game, which can be either the Producer or the Scrounger, a_t is the action chosen during the trial, $Q_t(s_t,a_t)$ denotes the value of the state-action pair at time t (commonly referred to as the Q-value), r_t denotes the reward received during the trial, and α represents the Learning Rate for updating the previously learned value. A higher value of α implies greater reliance on feedback from the current trial, while a value closer to zero incorporates learned experience from past interactions into the updated value.

After role assignment in each trial, the agent uses the Softmax Function (Equation 2) to decide between the two available actions corresponding to their role.

$$Pr(Q_i) = \frac{e^{Q_i/\tau}}{\sum_{j=1}^{A} e^{Q_j/\tau}} \tag{2}$$

In Equation 2, A denotes the number of possible actions for the agent (equal to 2 in this scenario), $Pr(Q_i)$ represents the probability that action i will be chosen, and the parameter τ in Equation 2 is known as the agent's temperature. τ modulates the trade-off between exploration and exploitation for the agent. Higher values of τ imply a more random decision-making process, while lower values (close to 0) indicate a preference for exploiting options with higher Q-values.

Throughout the simulation, both agents maintain a constant Learning Rate (α) and Temperature (τ) to eliminate confounding complexities in the interpretation of results and outcome states. This approach, referred to as an Action Learner, involves the agent learning the value of each action independently and considering only their own action, without regard to the opponent's action in the game.

In this study, we aim to investigate how different features of learning agents can affect the convergence of outcomes in the game. To accomplish this, we conducted tests of agent behavior using a 2×2 design. The first dimension concerns the agents' perspective on their payoff, referred to as selfish/utilitarian learning of payoffs. The second dimension concerns their ability to assign credit to their previous actions in the game, referred to as Non-temporal/Temporal learning in the game, which will be explained in the following paragraphs.

Accordingly, there are two distinct types of agents categorized by their approach to incorporating their opponent's payoff: 1) selfish agents, who only take into account their individual payoff in the game, and 2) utilitarian agents, who perceive the sum of both agents' payoffs as their payoff. To elaborate, in Equation 1, r_t for a selfish agent represents the assigned payoff based on their specific role, whereas, for a utilitarian agent, r_t is the sum of the payoffs in the converged outcome, regardless of their role. The objective behind implementing utilitarian agents is to explore how considering the opponent's outcome can influence the outcome state in the game.

In addition to changing the perspective on payoffs, in the learning paradigm described above, the agent operates as a Non-Temporal Learner. This implies that the agent does not maintain a historical record of its previous actions and updates its values based solely on the current state decision. This simplified learning framework does not take into account

any causal or temporal relationships that might influence the agent's decisions.

In contrast, we have developed an alternative approach that we call Temporal Learning. To better understand the intuition behind this method, imagine that you are faced with a situation in which your opponent is being moral with you. In this situation, you would give a bonus to previous actions to reach this state of social interaction. On the other hand, if your opponent is being immoral, you would devalue previous actions leading to an undesirable state. In the Temporal Learning method, the agent considers the history of its previous actions, specifically as the Scrounger, in determining the outcome of the current trial. Essentially, the agent tries to understand quasi-causally how its past actions as the Scrounger influence the current outcome in the game. Agents that follow this strategy are called Temporal Learners, and their learning mechanism is explained in Equation 3.

$$\begin{aligned} Q_t(s_{t-1} = \mathbb{P}_2, a_{t-1}) &= (1-\alpha).Q_{t-1}(s_{t-1} = \mathbb{P}_2, a_{t-1}) + \alpha.\gamma.r_t \\ Q_t(s_{t-2} = \mathbb{P}_2, a_{t-2}) &= (1-\alpha).Q_{t-1}(s_{t-2} = \mathbb{P}_2, a_{t-2}) + \alpha.\gamma^2.r_t \\ &\vdots \\ Q_t(s_{t-w} = \mathbb{P}_2, a_{t-w}) &= (1-\alpha).Q_{t-1}(s_{t-w} = \mathbb{P}_2, a_{t-w}) + \alpha.\gamma^w.r_t \end{aligned} \tag{3}$$

A Temporal Learner maintains a memory of their past actions as a Scrounger within a defined historical window, denoted by w. In addition, it updates the values associated with past state-action pairs using a discounted value of the current payoff, as shown in Equation 3. Here γ represents the discounting factor. At the end of each trial, a Temporal Learner first updates the relevant Q-value for the current state-action pair based on Equation 1. Subsequently, if the agent occupies the role of the Scrounger and is a Temporal Learner, it proceeds to update all corresponding Q-values for its previous Scrounger decisions with a discounted value of the current state's payoff ($\gamma^j.r_t$). This method allows temporal learners to consider the value of their past actions and optimize their long-term outcomes.

Results

Study 1: The impact of payoff calculation method on outcome change

In our first study, we examined the effect of agent characteristics on the outcome of the Producer-Scrounger Game. To do this, we ran 20,000 simulations of the game with different agent characteristics. Each simulation featured agents using a consistent set of parameters, with $\alpha = 0.1$ and $\tau = 1$ for both agents across all conditions.

The results of the simulations are shown in Figure 2. When both agents adopt a selfish stance, regardless of whether they are Temporal or Non-Temporal Learners, the Unproductive Outcome consistently emerges as the dominant equilibrium in the game. This finding demonstrates the robustness of the Unproductive Outcome in scenarios where self-interest is the guiding principle for both agents.

However, a notable shift occurs when both agents adopt a utilitarian perspective and consider each other's payoffs during the learning process. In these cases, the Utilitarian Outcome significantly outweighs other outcomes, highlighting the profound impact of pro-social consideration on the dynamics of decision-making.

Furthermore, our study examines a specific scenario where the Producer adopts a selfish perspective while the Scrounger adopts a utilitarian perspective. Interestingly, when both agents operate as Non-Temporal Learners, the Unproductive Outcome remains the dominant equilibrium, indicating that the utilitarian view does not significantly alter the converged state. Interestingly, a different pattern emerges when the agents are Temporal Learners. In this case, the utilitarian perspective shifts the outcome dynamics, leading to the dominance of the Private Property Outcome. However, when the opposite scenario occurs, that is, when the Producer adopts a utilitarian perspective and the Scrounger becomes selfish, the Utilitarian Outcome becomes extremely dominant in the game.

Study 2: Strategy Learning vs. Action Learning

Consistent with the approach outlined in the Methods section, our agents operate primarily as Action Learners, where their decision-making process revolves around the evaluation of action values. However, the second study of our research introduces an alternative approach known as Strategy Learners.

A strategy defines an agent's response based on their assigned role in the game. For example, as shown in Figure 3-a, a strategy might include choosing "NP" as the Producer and "T" as the Scrounger, or perhaps "P" as the Producer and "T" as the Scrounger. Strategy Learners, as the name implies, evaluate and learn the value associated with different strategies in the game, as opposed to focusing solely on action values.

To implement this approach, as shown in Figure 3-a, each agent maintains a Q-value for each possible strategy within the game. During each trial, the agents use the Softmax Function (Equation 2) to make their strategy-based choices from the array of possible strategies. Depending on their role in the game, they then determine their actions for that particular trial. For example, if Q_3 represents the Softmax outcome and the agent assumes the role of Scrounger, the agent's decision is "L", whereas if they act as Producer, their choice is "P". Following the decision process and subsequent feedback from the environment, the agent updates the Q-value of the strategy actually employed in the game, not the strategy they initially decided upon. For example, if Q_3 serves as a Softmax for an agent, and the agent chooses "P", the opponent's choice of "T" causes an update to Q_4. This update concerns the value associated with 'P' for the Producer and "T" for the Scrounger. In essence, Strategy Learning differs from Action Learning in that both agents gain access to their opponent's action (not payoff), make decisions based on their respective strategies, and subsequently update the played strategy.

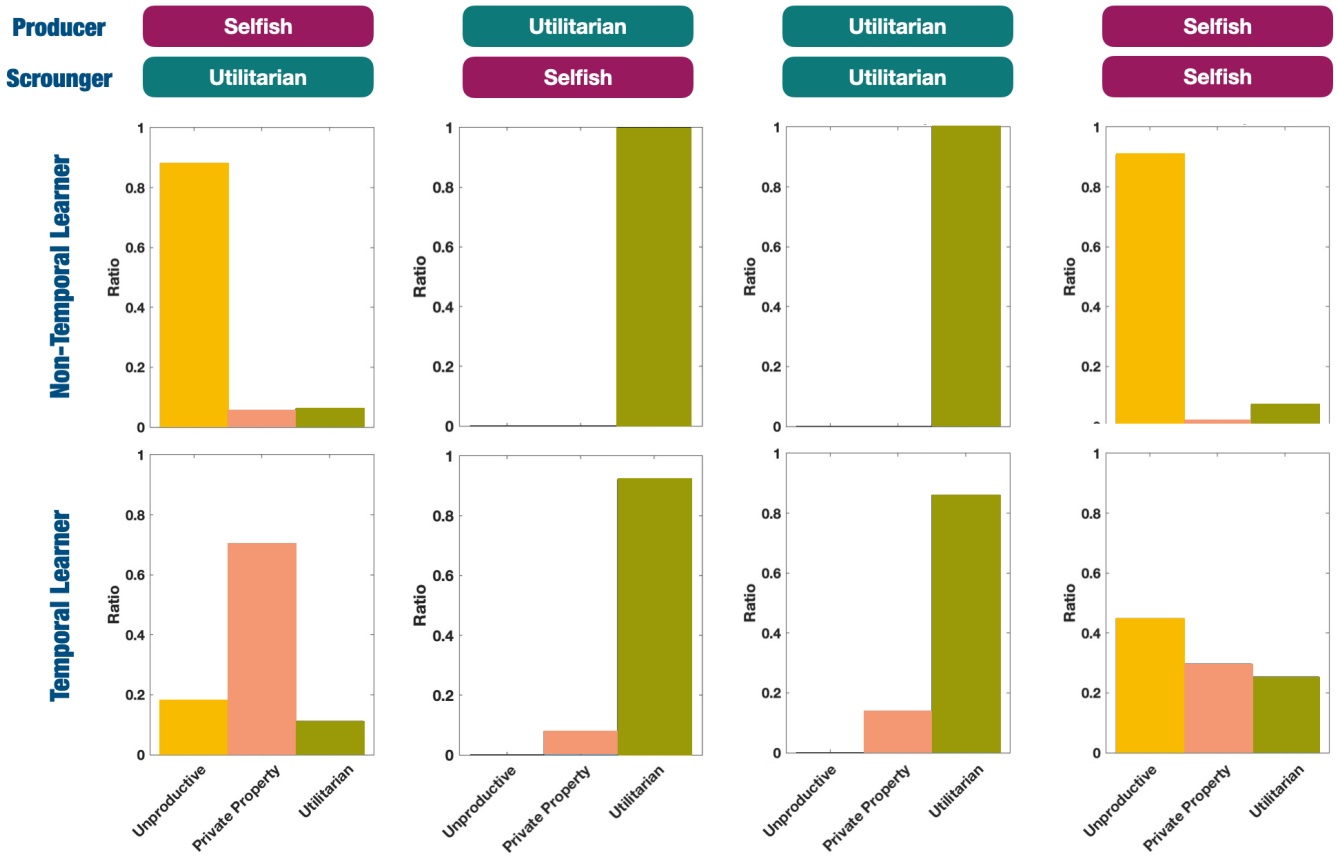

Figure 2: The impact of Temporal/Non-Temporal Learning on the converged outcome considering different agent characteristics. The ratio of each outcome changes during the Non-Temporal (top) and Temporal Learning (bottom) of the game.

To examine how Strategy Learners navigate the Producer-Scrounger Game, we created a modified version of the game in which the Utilitarian Outcome is no longer aligned with social welfare maximization due to a change in the Scrounger's payoff (Figure 3-b-right). The purpose of this manipulation was to measure the adaptability of strategy learners in response to variations in the pursuit of social welfare maximization.

Figure 3-b shows the results of 20,000 simulations under different conditions. In all simulations, the agents assumed a selfish disposition with fixed parameters of $\alpha = 0.1$ and $\tau = 1$ - as in Study 1. As expected, and consistent with the results of Study 1, when agents operate as selfish Action Learners, the Unproductive Outcome remains the dominant equilibrium, and the pursuit of social welfare does not significantly affect the converged state. However, in the basic version of the game (left plots), when selfish Agents adopt the Strategy Learning approach, they converge to the Utilitarian Outcome, which maximizes the sum of payoffs. Notably, when the Private Property Outcome emerges as the social welfare maximizer (right plots), Strategy Learners converge to this state.

These results suggest that even without considering the payoffs of their opponents (like utilitarian agents) and without engaging in temporal value learning, agents can reach collectively optimal outcomes by adapting strategies. This provides valuable insights into potential models of human cooperation and their implementation in artificial agents for everyday interactions, paving the way for more nuanced and effective strategies in various applications.

Discussion

In Study 1, we examined how agent characteristics affect the outcome dynamics of the Producer-Scrounger game. Our exploration of specific scenarios, particularly those in which the Producer adopts a selfish stance while the Scrounger adopts a utilitarian perspective, revealed intriguing dynamics. Specifically, in the absence of Temporal Learning, the Unproductive Outcome asserted persistent dominance. However, when agents engaged in Temporal Learning, a remarkable shift occurred, resulting in the dominance of the Private Property Outcome. In Study 2, we introduced Strategy Learners as a novel paradigm that differs from conventional Action Learners in that it focuses on the evaluation and acquisition of strategies rather than individual action values. Action Learners, driven by the imperative of payoff maximization, consistently converged on the Unproductive Outcome. In contrast, Strategy Learners exhibited a remarkable degree of adaptability, converging on either the Utilitarian or Private Property

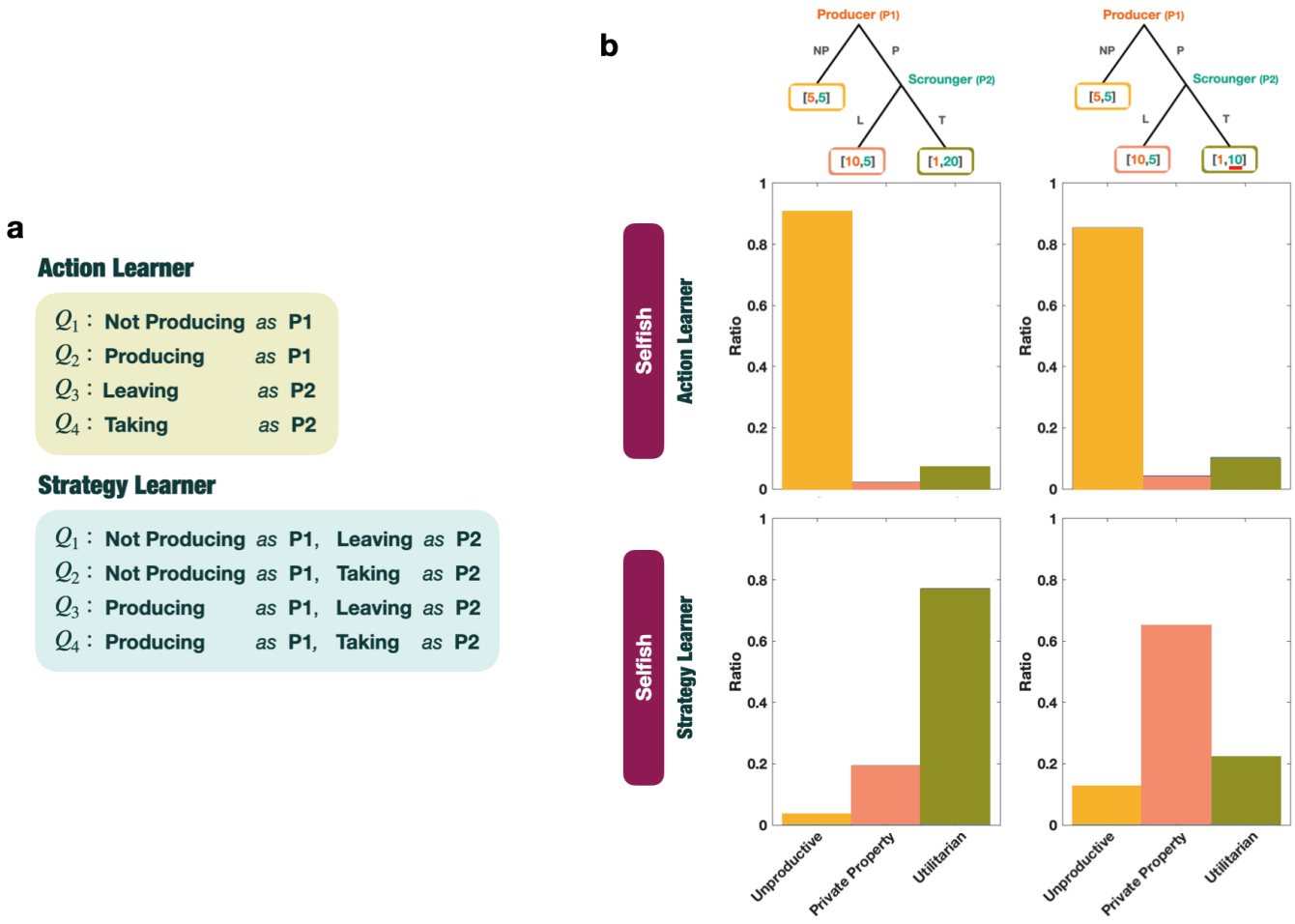

Figure 3: Action vs. Strategy Learning conditions in two different versions of the Producer-Scrounger Game. (a) An Action Learner learns q-values for each action, whereas a Strategy Learner assigns q-values for each strategy profile. (b) Outcome ratios in selfish agents using Action Learning (top) vs. Strategy Learning (bottom).

Outcome depending on the prevailing social welfare maximizer. Our computational simulations provide a piece of evidence that even value-based consequentialist RL agents, which consider the temporal value of their actions, can converge on outcomes that respect private property principles rather than strictly adhering to utilitarian considerations. This underscores the profound influence of considering the payoffs of other players, where the calculation of temporal value plays a central role in determining the game's convergence point.

Furthermore, the strategy learning paradigm introduced in this study avoids direct consideration of the other player's payoff. Instead, it only observes the actions of the other player and updates the value of the strategy that is actively implemented in the game. This novel approach implies that agents do not evaluate their individual actions, but rather the collective state they achieve together. Importantly, the implicit understanding that societal roles can change contributes to convergence towards outcomes that maximize social welfare. In line with our study, Kuzmics et al. (2014) explore the implications of symmetric play in repeated allocation games, providing insight into the influence of players' continuation payoffs on equilibrium outcomes. This suggests that anticipation of other players' payoffs leads to adjustments in strategy choice to achieve favorable outcomes. Moreover, our approach assumes that individuals recognize the collective nature of decision-making and thus evaluate the strategy that culminates in the game rather than their individual actions. This is consistent with the inherent limitation of not having direct access to the utility of others; consequently, individuals use themselves as a proxy to evaluate the payoffs and future motivations of their counterparts.

As we move toward a future in which AI systems are increasingly confronted with complex decision spaces, understanding the complications of human moral reasoning becomes imperative. By recognizing the impact of agent characteristics, learning strategies, and consideration of others' payoffs in our computational models, we provide insights for developing AI systems that can emulate and understand human-like moral decision-making.

References

André, J.-B., Fitouchi, L., Debove, S., & Baumard, N. (2022). An evolutionary contractualist theory of morality. *PsyArXiv*.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122–142.

Christian, B. (2021). *The alignment problem: How can machines learn human values?* Atlantic Books.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273–292.

Ecoffet, A., & Lehman, J. (2021). Reinforcement learning under moral uncertainty. In *International conference on machine learning* (pp. 2926–2936).

Frey, B. S., & Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic literature*, *40*(2), 402–435.

Gauthier, D. (1987). *Morals by agreement*. clarendon Press.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & language*, *22*(1), 1–21.

Herlau, T. (2022). Moral reinforcement learning using actual causation. In *2022 2nd international conference on computer, control and robotics (icccr)* (pp. 179–185).

Horita, Y., Takezawa, M., Inukai, K., Kita, T., & Masuda, N. (2017). Reinforcement learning accounts for moody conditional cooperation behavior: experimental results. *Scientific reports*, *7*(1), 39275.

Kuzmics, C., Palfrey, T., & Rogers, B. W. (2014). Symmetric play in repeated allocation games. *Journal of Economic Theory*, *154*, 25–67.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169.

Nguyen, D., Venkatesh, S., Nguyen, P., & Tran, T. (2020). Theory of mind with guilt aversion facilitates cooperative reinforcement learning. In *Asian conference on machine learning* (pp. 33–48).

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, *304*(5669), 452–454.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045.

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, *2*, 64–69.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Selten, R., & Bielefeld, R. S. (1988). *Reexamination of the perfectness concept for equilibrium points in extensive games*. Springer.

Shadmehr, R., Reppert, T. R., Summerside, E. M., Yoon, T., & Ahmed, A. A. (2019). Movement vigor as a reflection of subjective economic utility. *Trends in neurosciences*, *42*(5), 323–336.

Sorensen, T., Jiang, L., Hwang, J., Levine, S., Pyatkin, V., West, P., ... others (2023). Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *arXiv preprint arXiv:2309.00779*.

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 1–7.

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

Yudkowsky, E. (2016). The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, *4*.