Reinforcement Learning, Social Value Orientation, and Decision Making: Computational Models and Empirical Validation

Peter Duggins (psipeter@gmail.com)

Chris Eliasmith (celiasmith@uwaterloo.ca)

Terrence C. Stewart (terrence.stewart@nrc-cnrc.gc.ca)

Computational Neuroscience Research Group
Department of Systems Design Engineering, University of Waterloo

Abstract

Social environments often impose tradeoffs between pursuing personal goals and maintaining a favorable reputation. We studied how individuals navigate these tradeoffs using Reinforcement Learning (RL), paying particular attention to the role of social value orientation (SVO). We had human participants play an interated Trust Game against various software opponents and analyzed the behaviors. We then incorporated RL into two cognitive models, trained these RL agents against the same software opponents, and performed similar analyses. Our results show that the RL agents reproduce many interesting features in the human data, such as the dynamics of convergence during learning and the tendency to defect once reciprocation becomes impossible. We also endowed some of our agents with SVO by incorporating terms for altruism and inequality aversion into their reward functions. These prosocial agents differed from proself agents in ways that resembled the differences between prosocial and proself participants. This suggests that RL is a useful framework for understanding how people use feedback to make social decisions.

Keywords: Reinforcement Learning; Trust Game; Instance-Based Learning; Semantic Pointer Architecture; Altruism; Inequality Aversion

Introduction

In social decision making, individuals must gather information, weigh alternatives, and select actions in environments that contain other intelligent agents. Such environments often involve tradeoffs between short- and long-term rewards and between individual and collective interests. Researchers in psychology, economics, and neuroscience study these tradeoffs, and the behaviors they encourage, using social dilemmas such as the Prisoner's Dilemma, the Ultimatum Game, and the Trust Game. To navigate these social dilemmas successfully, players must build mental models of their social environments and adapt their behavior in response to decisions made by other individuals. We are interested in the cognitive mechanisms that underlie such learning, particularly the role of social value orientation (SVO) in promoting cooperative behavior and achieving long-term collective outcomes.

In this paper, we investigate how humans and simulated agents learn to play the Trust Game (TG), studying both the distributions and dynamics of behaviors as heterogeneous individuals explore and settle on strategies against various types of opponents. We adopt a Reinforcement Learning (RL) approach to describe the learning process and propose that SVO critically influences learned behaviors. We operationalize SVO within the RL framework, implement RL in two different cognitive architectures, and compare simulated data from heterogeneous populations of these agents to human data from the TG, noting in particular the effects of SVO on behavior in different social settings. Our results are consistent with theoretical and empirical accounts of the relationship between learning, SVO, and behavior, and demonstrate that SVO can be incorporated into a computational theory of learning that generalizes across cognitive architectures.

Background

The Trust Game (TG) is two-player, turn-based game in which individuals repeatedly receive and reallocate resources in a sequential manner. In each turn of the game, the first player (the *investor*), receives ten coins, then gives some of these coins to the second player (the *trustee*), keeping the rest. The *trustee* receives three times this many coins. Finally, the *trustee* returns some number of the resulting coins to the *investor*. A single game consists of five turns. Each player's final score is the total number of coins collected across all five turns. In the TG, greater rewards are earned if both players invest and return generously, but each player will only do so if they trust their opponent to reciprocate in future rounds.

Behavior in the TG has a clear prosocial component that is distinct from maximizing individual rewards: while *investor* behavior is most closely correlated with expectations of repayment and perceived trustworthiness, *trustee* behavior is most closely correlated with prosocial tendencies (Ashraf et al., 2006). For instance, in one-turn versions of the TG, many *trustee*s will return coins even though they cannot receive returns from future interactions. Prosocial behavior is motivated by numerous drives, including reciprocity, inequality-aversion, and altruism (Pletzer et al., 2018; Declerck et al., 2013), and the activity of numerous brain structures correlates with prosocial values estimates and prosocial behaviors, including the amygdala, striatum, TPJ, mPFC, and dlPFC (Haruno & Frith, 2010; Hutcherson et al., 2015). Evidence from psychology and neuroscience supports the idea that learning is a critical component in social dilemmas: individuals adapt their strategies in response to specific instances of betrayal (Lount Jr et al., 2008), to the opponent's behavior in the recent past (Engle-Warnick & Slonim, 2004), and to trust estimates made in previous games (Collins et al., 2016).

Reinforcement Learning (RL) is a widely-acknowledged framework for understanding how humans and other animals update their behavior based on external feedback (Sutton &

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Barto, 2018): numerous studies have shown that the signals and learning rules proposed by RL map onto reward-prediction errors and synaptic changes in the brain (Glimcher, 2011). Furthermore, recent frameworks such as RLDM have begun to lay out the relationship between RL, decision making (DM), and social value orientation (SVO), noting how various prosocial motivations may arise from different forms of learning (e.g., model-based, model-free, and associative) at various levels of abstraction, in different tasks, and through different brain regions (Gesiarz & Crockett, 2015). However, more computational work is needed to elucidate the relationship between RL algorithms and emergent prosocial behavior; specifically, the field needs more models that incorporate SVO into learning rules and value functions in a manner that (a) is cognitively and neurally plausible, (b) can be generalized to multiple cognitive architectures and behavioral tasks, and (c) explains a variety of empirical results.

Here, we use computational models to investigate the relationship between learning, SVO, and prosocial behavior. To model SVO, we incorporate two additional term into the reward function of RL agents; this encourages agents to consider the rewards obtained by other individuals in social dilemmas. In keeping with other computational models of SVO (Hutcherson et al., 2015; Collins & Juvina, 2021; McKee et al., 2020), an agent's overall reward is a weighted combination of self-reward, other-reward, and reward-inequality, where the relative weighting determines the agent's degree of SVO. Using this reward function, we train two distinct cognitive architectures to play the TG. Our first architecture is a Deep Q-Network (DQN), a neural network trained with backpropagation that has previously been used to learn complex, human-like behavior in multiplayer games (Wang et al., 2018). Our second architecture is based on ACT-R (Anderson et al., 2004), an integrated theory of cognition supported by an extensive history of cognitive models validated by human behavioral data. In this architecture, the retrieval and utilization of episodic memories is governed by instance-based learning (IBL) and blended retrieval (Thomson et al., 2015), two mechanisms that are closely aligned with observed human recall in natural settings and decision making tasks (Gonzalez et al., 2003) and have been used as the basis for decision making in simulated agents playing two-player social games (Lebiere et al., 2009). For each architecture, we simulate and train a heterogeneous population of agents, then compare the distribution and dynamics of simulated data with human data in the TG. We then discuss whether RL, and specifically our implementation of SVO, is a suitable framework for studying prosocial decision making, independent of the computational implementation of agent's internal model.

Methods

To model human learning in the TG, we used RL to train agents from both different computational architectures. Agents learn a "Q-function" which assigns a value to (state, action) pairs, then selects the action a with the highest estimated value when present in state s. The state consists of the current turn (1 through 5) and the number coins available (0 to 30). Actions determine the number of coins the agents give and keep. The Q-function is updated using the standard TD(0) Q-learning formula,

$$\Delta Q(s,a) = \alpha\big[r + \gamma \max_{a^*} Q(s',a^*) - Q(s,a))\big], \qquad (1)$$

where α is the learning rate, r is the reward, γ is the discount factor, and $\max_{a^*} Q(s',a^*)$ is the maximum over possible actions in the next state s'. In order to explore the (state, action) space, agents take random actions with a probability ε, which decays over the course of learning. If a random action is not taken, the agent chooses the action with the greatest Q-value. To model SVO, agents compute a weighted sum between the rewards they themselves received, the rewards their opponent received, and any inequalities between them:

$$r = w_s r_s + w_o r_o - w_i |r_s - r_o|, \qquad (2)$$

where w_s, w_o, and w_i are the weighs for self-reward, other-reward, and inequality, and the rewards r_s and r_o are the coins earned on any given turn. For convenience, we fix $w_s = 1$ and interpret w_o and w_i as an agent's SVO. When initializing heterogeneous agent populations, each individual is given a unique w_o and w_i, a discount factor γ, and a learning rate α.

Although Q-learning, exploration, and SVO are common features of all our agents, the two architectures implement learning, function approximation, and memory in different ways (Figure 1). Our Deep Q-Network agent is a three-layer network with rectified linear activation, a loss function of ΔQ^2, and Adam backpropagation. To facilitate comparisons between the learning trajectories of DQN agents and our other agents, we did not use an experience replay buffer.

Our Instance-Based Learning agent contains an episodic memory and a working memory, which work together via queries and retrievals to recall relevant pieces of information from previous experiences. These "chunks" contain the state s, the selected action a, the returned reward r, and the estimated value Q. When choosing an action, the agent looks through all chunks in episodic memory and loads into working memory those chunks which satisfy two criteria: the chunk has sufficient activation due to recent or frequent use, and its state is sufficiently similar to the current state. The activation of a chunk is given by the standard ACT-R equations, and state similarity depends on the number of coins available (normalized absolute difference) and the turn number (zero if turn number is different, otherwise one). The value of each potential action is calculated using blended retrieval,

$$\hat{Q}(a) = \sum_{i}^{M} Q_i(a) A_i, \qquad (3)$$

that is, all chunks i where action a was taken are queried for their estimated value Q, weighted by their activation A, and summed. Finally, the Q value assigned to each new chunk

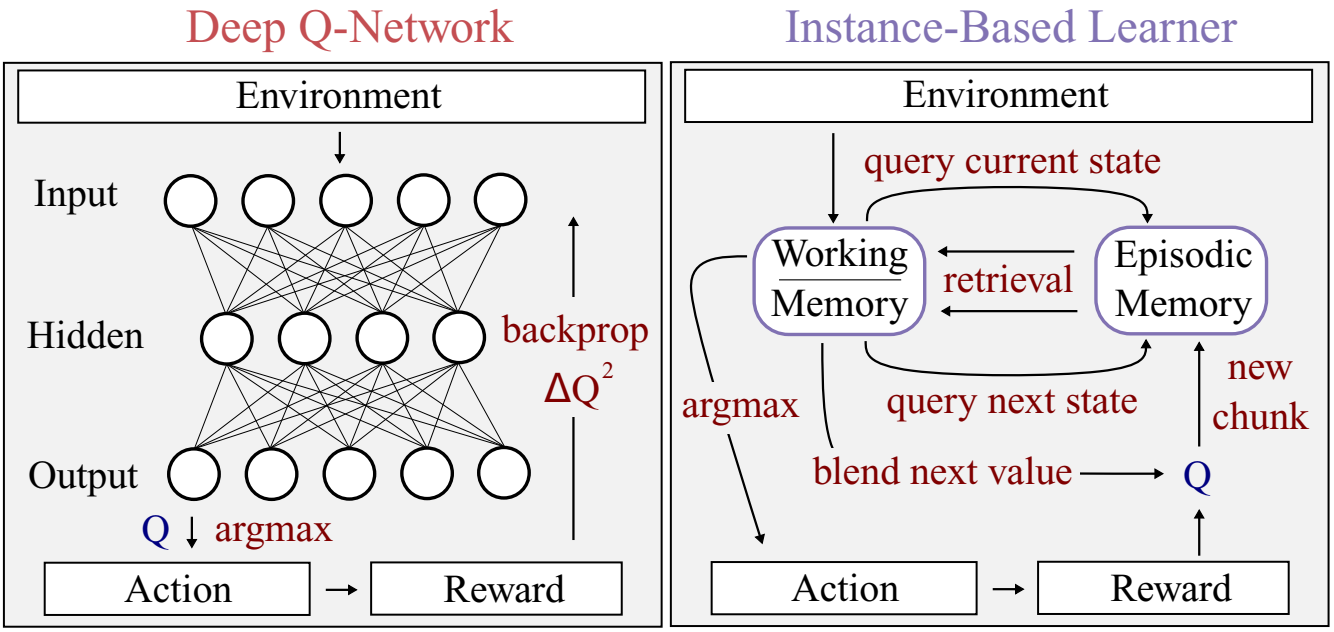

Figure 1: **Network Architectures for the RL agents**. Boxes represent inputs/outputs, circles represent individual neurons, and purple ovals represent ACT-R memory systems.

is equal to the reward returned for that action plus the discounted expected value of "future" chunks. For this expectation, the agent recalls all chunks j with sufficient similarity to the game's *next* state (s_{t+1}) and blends their values

$$Q_i = r + \gamma \sum_j^M Q_j A_j, \tag{4}$$

where the sum is over the chunks j that pass the activation and state similarity thresholds for s_{t+1}. This mechanism is a novel realization of RL within the ACT-R framework, which typically learns the value of procedural rules (actions) directly.

To validate the simulated data produced by these agents, we also ran a simple human experiment through Amazon Mechanical Turk in which participants learned to play the TG against simulated opponents. Participants completed a tutorial that introduced the rules and strategy of the game, then played thirty five-turn games alternating between the *investor* and the *trustee*. The *investor* began each round with ten coins, and the transfer sent to the *trustee* was tripled. Participants earned \$0.10 per game plus \$0.003 per coin they collected, incentivizing them to play strategically rather than quickly. Participants were classified as "proself" or "prosocial" based on a post-trial survey:

- I tried to earn as many points for myself as possible, without considering my opponent's score. (N=83)
- I tried to achieve a high score for both myself and my opponent. (N=115)

Participants were secretly sorted into two groups: the first group faced opponents who could profitably be exploited with a greedy strategy, while the other group faced opponents where the best strategy was to be consistently generous. To minimize the antisocial effects that humans exhibit when playing against simulated opponents with fixed strategies (Mota et al., 2016), we strove to make our agents human-like: all agents (a) played according to an adaptive "Tit-for-Tat" (T4T) strategy, responding generously if the human was generous and greedily if the human was greedy, (b) had randomized response times that matched human response times, and (c) were initialized with parameters that controlled the initial response and the magnitude of the T4T update, producing (i) heterogeneous behaviors across turns and games and (ii) exploitable versus cooperative behaviors in the two experimental groups. We were interested in whether participants would learn high-reward strategies against the opponents they faced, whether their social orientation would influence this process, and to what degree our RL agents would reproduce the dynamics and distributions of these behaviors.

Results

We began by training our RL agents against simple software opponents, tuning model hyperparameters until agents learned the optimal policy. We then created heterogeneous populations of agents by modestly varying these hyperparameters (notably γ and α) and by introducing SVO (w_o and w_i), which we set to zero for proself agents and to random values between 0 and 0.5 for prosocial agents. We trained these populations against the same T4T opponents used in the human experiment. To compare the human and agent data, we transformed the actions taken by individuals into a normalized *generosity*, which indicates the fraction of available coins the

individual transfers on each turn.

To examine the final strategies learned by humans and agents, we plot the distribution of generosities in the final $n = 3$ games[1] versus different opponents, when playing as different players, and grouped by SVO, Fig. 2. Looking at the human data, we observe several patterns that are consistent with the RLDM literature and demonstrate that this dataset is a valid point of comparison for our RL agent data. First, human behaviors are diverse: generosities vary between zero (keeping all available coins) and one (sending all available coins) in every condition, indicating that participants learn a wide variety of strategies that differ across turns and/or between individuals. Second, the distribution of generosities differs significantly (a) when playing against different opponents and (b) with participant SVO. To quantify differences between generosity distributions, we use the two-sample Kolmogorov-Smirnov test: the magnitude of the test statistic indicates the difference between these distributions, while the p-value describes whether this difference is statistically significant. Participants adopted significantly different behaviors against each opponent ($p < 0.0001$ in 4/4 conditions), confirming that humans learn strategies that are adapted to the social environment; and prosocial participants learned significantly different behaviors than did proself participants ($p < 0.0001$ in 3/4 conditions). Furthermore, prosocial individuals were significantly more generous (Welch's t-test, $t = 17.1$, $p < 0.0001$, $\Delta\bar{G} = 0.15$) and significantly less likely to defect ($generosity < 0.2$) on the final turn when playing the as the *trustee* ($t = 2.8$, $p = 0.007$, $\Delta\bar{P} = 0.16$); these results confirm that SVO influences human DM with regards to both the learned distributions of generosities and several descriptive metrics.

We then compared the simulated data from our RL agents to the human data. In many cases, RL agent strategies captured interesting features of the human data: for instance, the proself IBL agents have a tendency to be less generous on each successive turn when playing the *investor* against a *generous* opponent, an unexpected (suboptimal) pattern that also appears in the corresponding human data. A majority of agents also learned to defect in the final turn when playing as the trustee, but there remained a significant percentage of both humans and agent that did not discover (or did not adopt) this strategy. In other instances, the strategies learned by RL agents differed significantly from the human data: for example, many agents architectures returned more than 50% of the available coins when playing as the *trustee*, a behavior that was rarely exhibited by humans.

With respect to SVO, significant differences between the behaviors of proself and prosocial agents were observed in all conditions (4/4 conditions for both agents, $p < 0.0001$). Prosocial agents were more generous on average than their proself counterparts ($\Delta\bar{G} = 0.08$ (DQN) and 0.03 (IBL), both $p < 0.0001$), and were also less likely to defect on the final turn when playing the *trustee* ($\Delta\bar{P} = 0.34$ (DQN) and 0.22 (IBL) both $p < 0.0001$). However, the learned behaviors of prosocial agents were *not* significantly better fits to the prosocial human data than were the learned behaviors of proself agents (as measured by the KS-test).

We also analyzed how behaviors changed during the course of learning for participants and our RL agents. To examine how individual's strategies converged with experience, we plotted the similarity between an individual's generosity distribution in each game and their generosity distribution in the final $n = 3$ games, Fig. 3. These plots show qualitatively different patterns of convergence between agent architectures, especially in reference to the human data. Participant strategies converged in a fairly linear manner, although in some conditions there was a period of rapid change in either the initial or final games. Additionally, there was less within-individual variability in the final strategies of *investors* than there was for *trustees*. IBL strategies converged in a similar manner, exhibiting linear change in most conditions and more variation among endgame *trustee* strategies. In contrast. DQN strategies tended to converge only during the final games. There were no noticeable differences between the learning trajectories of proself and prosocial humans or agents with respect to patterns of convergence.

Discussion

In this paper, we investigated how humans learn to make social decisions by observing the development of their strategies in the Trust Game and modelling this adaptation using Reinforcement Learning. As expected, our empirical data showed that participants learned a variety of strategies that differed significantly between individual, were well-suited to the opponents they faced, and were strongly influenced by SVO. To model the learning process and account for the effects of SVO, we designed and trained two classes of agents from modern cognitive architectures, then endowed them with SVO by adding terms for altruism and inequality-aversion into their reward functions. These agents learned effective TG strategies that captured several features of the human data, including the tendency to defect when future reciprocation was impossible, sustained generosity by prosocial individuals against greedy opponents, and a steady convergence of behaviors towards a final strategy as exploration gave way to exploitation.

The observed differences in learning and behavior between agent architectures, and in contrast to the human data, suggest several tentative conclusions. With regards to learning, humans develop coherent strategies much faster than RL agents (15 versus 200-400 games). While our data suggest that some RL agents converge on human-like strategies according to human-like dynamics, humans probably utilize additional cognitive mechanisms to speed up learning and decrease the experience needed to discover effective strategies. For in-

[1]We chose $n > 1$ to (a) increase the amount of data and give greater statistical power to our tests, and (b) to ensure that random fluctuations in the behavior of opponent agents (from game to game) did not skew our analysis of final strategies. The reported results remain consistent for various choices of n.

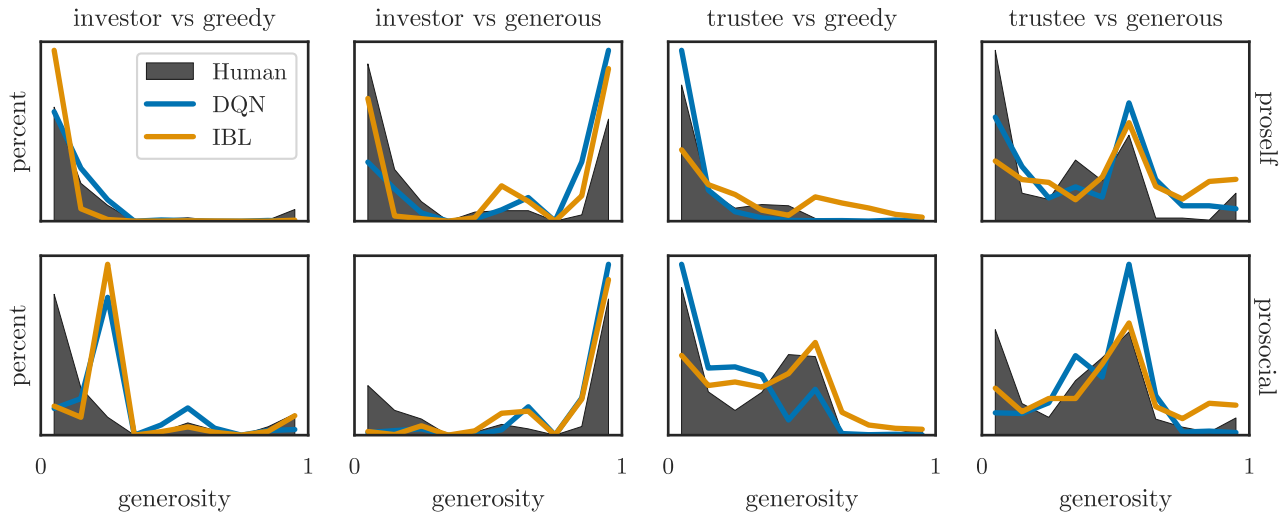

Figure 2: **Strategies learned by humans and RL agents, divided according to player, opponent identity, and SVO**. The y-axis of these histograms indicates the percent responses in each generosity bin, with a polygon interpolation applied for visualization. Each column represents data from one experimental condition (player and opponent), while the top and bottom rows represent proself and prosocial individuals, respectively.

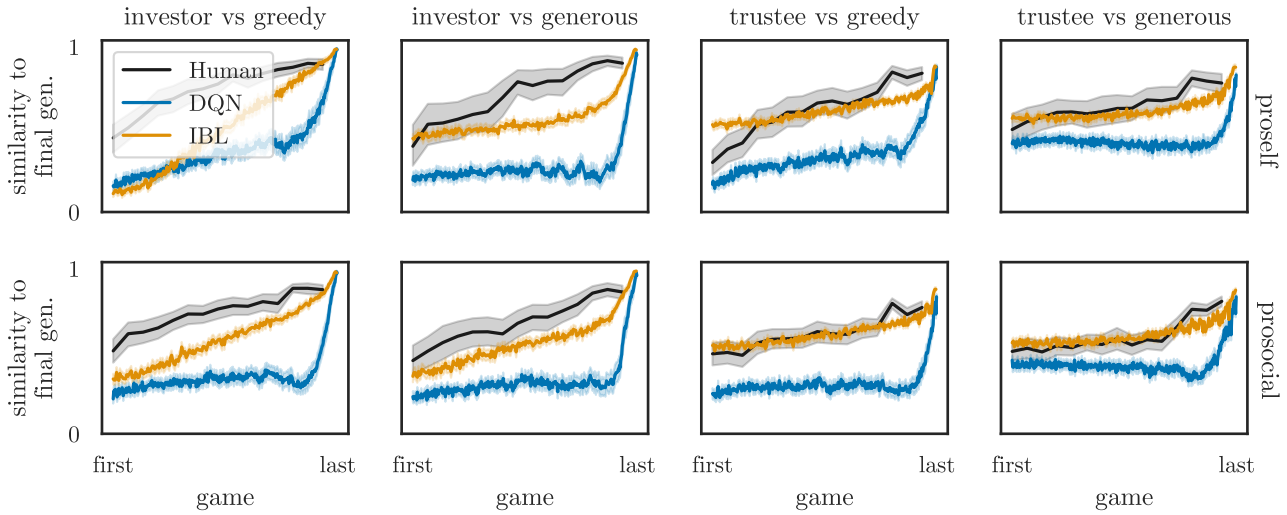

Figure 3: **Convergence of behaviors for humans and RL agents**. The x-axis represents cumulative experience playing the TG, while the y-axis plots the similarity between the generosity distribution in the current game and the generosity distribution in the final $n = 3$ games. Shaded regions represent confidence intervals across individuals. Note that the x-axis is scale-free: it does not indicate the number of games required to learn a strategy.

stance, where our RL models rely on random exploration to explore the (state, action) space of the TG, humans probably use previously-acquired knowledge and heuristics to facilitate efficient learning. With regards to SVO, we found that prosocial agents learned behaviors that were statistically-distinct from their proself counterparts, and observed that prosocial agents made more generous transfers and defected less often than proself agents, all trends that we also noticed in the human data. However, when comparing the final distribution of agent generosities to human generosities in each experimental condition, we found that prosocial agents were not significantly better fits than their proself counterparts. From these contrasting results, we conclude that our operationalization of SVO in RL agents was sufficient to reproduce several important qualitative trends in human behavior, but insufficient to reproduce the exact strategic differences adopted by prosocial humans. It is important to note, however, that we did not optimize model parameters to fit the human data: we simply created heterogeneous populations of agents with w_o and w_i drawn from a wide range of possible values, and compared the resulting distribution of learned generosities to the learned generosities of humans. In future work, we plan to optimize these parameters to fit individual participants, then investigate whether the optimal SVO parameters from our agents predict participants' SVOs and improve the overall fit of our RL models.

Experimental conditions in which prosocial agents poorly fit the prosocial human data often reflect understandable failures of our SVO mechanism. Specifically, agents with nonzero altruism and inequality aversion continued investing against greedy opponents, a trend that was also apparent in prosocial humans. Participants manifested this trend by occasionally investing the maximum amount, but quickly returned to zero-investment strategies after observing a lack of reciprocation. Our RL agents, on the other hand, manifested this generosity as steady levels of small investment, a strategy which satisfied the agent's concern for others' rewards but ended up scoring lower on our similarity metric than the invariant zero-investment strategy of proself agents. In future work, we would like to explore the relationship between reciprocity and sustained cooperation, a trend observed in prosocial humans (Pletzer et al., 2018) that has been operationalized in other computational models of RL using IBL (Juvina et al., 2015) and which may relate to context-dependent weighting of proself and prosocial value in vmPFC (Declerck et al., 2013).

Agent architecture made a noticeable impact on the dynamics and distributions of learned generosities, despite being governed by identical update rules (Eq. 1) and exploration schedules. While DQN agents learned strategies that led to high average rewards in the TG, they were some of the poorest fits to human data: these agents required twice as much training data, demonstrated little convergence in the early games, and learned policies with few minimal turn-to-turn variation. Given that this architecture has fewer cognitive and biological constrains, it is not surprising that it gave poorer matches to human data (but recall that we did not train the network with this objective in mind). IBL agents, on the other hand, successfully captured the dynamics of convergence and turn-to-turn variation. This correspondence may reflect the cognitively-grounded mechanisms for episodic memory formation and recall in the architecture. However, the differences between proself and prosocial IBL agents were less apparent than they were in the human data, and IBL agents playing the trustee behaved more erratically than did our human participants. We also built and trained a third class of agent based on the Semantic Pointer Architecture (SPA), a framework for building biologically-constrained neural networks that perform cognitive tasks and reproduce neural and behavioral data (Eliasmith, 2013). In this model, we used an online, error-driven learning rule to implement Q-learning within the network, as well as a short-term memory system to recall previous states and an independent-accumulator model for action selection that resembles the drift-diffusion model (Ratcliff & McKoon, 2008). This agent also successfully learned to play the TG, and preliminary results showed that it behaved similarly to the IBL agent, with respect to the dynamics of convergence and the differences between proself and prosocial agents. Unfortunately, due to time constraints, we were not able to gather sufficient data to include these agents in the above analyses, but future work will continue the development of these neurally-plausible agents.

Our results suggest that RL is a sensible framework for modelling the learning process behind social decisions and can be implemented in various cognitive architectures. They also show that SVO can be operationalized into the RL framework if individuals consider the rewards of others when estimating the value of states and actions. However, given the complexity of empathy and mentalizing in the human brain, and the inability of our prosocial agents to reproduce specific prosocial behaviors, more work is needed to extend these learning mechanisms within cognitively plausible architectures.

References

Anderson, J. R., et al. (2004). An integrated theory of the mind. *Psychological review*, *111*(4).

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Exp. Econ.*, *9*(3).

Collins, M. G., & Juvina, I. (2021). Trust miscalibration is sometimes necessary: An empirical study and a computational model. *Frontiers in Psychology*, *12*.

Collins, M. G., Juvina, I., & Gluck, K. A. (2016). Cognitive model of trust dynamics predicts human behavior within and between two games of strategic interaction with computerized confederate agents. *Front. Psychol*, *7*.

Declerck, C. H., Boone, C., & Emonds, G. (2013). When do people cooperate? the neuroeconomics of prosocial decision making. *Brain and cognition*, *81*(1).

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Engle-Warnick, J., & Slonim, R. (2004). The evolution of strategies in a repeated trust game. *J Econ Behav Organ*.

Gesiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and pavlovian prosocial behavior. *Frontiers in behavioral neuroscience*, *9*, 135.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *PNAS*, *108*(Supplement 3).

Gonzalez, C., Lerch, J., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci*, *27*(4).

Haruno, M., & Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature neuroscience*, *13*(2), 160–161.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451–462.

Juvina, I., Lebiere, C., & Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of applied research in memory and cognition*, *4*(3), 197–211.

Lebiere, C., Stewart, T., & West, R. (2009). Applying cognitive architectures to decision-making. In *Cogsci* (Vol. 31).

Lount Jr, R., Zhong, C.-B., Sivanathan, N., & Murnighan, J. (2008). Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Pers Soc Psychol Bull*.

McKee, K. R., Gemp, I., McWilliams, B., Duéñez-Guzmán, E. A., Hughes, E., & Leibo, J. Z. (2020). Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*.

Mota, R., et al. (2016). Playing the 'trust game'with robots: Social strategies and experiences. In *Ro-man 2016.*

Pletzer, J. L., Balliet, D., Joireman, J., Kuhlman, D. M., Voelpel, S. C., & Van Lange, P. A. (2018). Social value orientation, expectations, and cooperation in social dilemmas: A meta–analysis. *European Journal of Personality*, *32*(1), 62–83.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

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

Thomson, R., et al. (2015). A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *J. Appl. Res. Mem*, *4*(3).

Wang, W., Hao, J., Wang, Y., & Taylor, M. (2018). Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv:1803.00162*.