

## Warmth and competence in human-agent cooperation

Kevin R. McKee  · Xuechunzi Bai  ·  
Susan T. Fiske 

Received: 15 Feb 2023 / Accepted: 12 Apr 2024

**Abstract** Interaction and cooperation with humans are overarching aspirations of artificial intelligence (AI) research. Recent studies demonstrate that AI agents trained with deep reinforcement learning are capable of collaborating with humans. These studies primarily evaluate human compatibility through “objective” metrics such as task performance, obscuring potential variation in the levels of trust and subjective preference that different agents garner. To better understand the factors shaping subjective preferences in human-agent cooperation, we train deep reinforcement learning agents in Coins, a two-player social dilemma. We recruit  $N = 501$  participants for a human-agent cooperation study and measure their impressions of the agents they encounter. Participants’ perceptions of warmth and competence predict their stated preferences for different agents, above and beyond objective performance metrics. Drawing inspiration from social science and biology research, we subsequently implement a new “partner choice” framework to elicit *revealed* preferences: after playing an episode with an agent, participants are asked whether they would like to play the next episode with the same agent or to play alone. As with stated preferences, social perception better predicts participants’ revealed preferences than does objective performance. Given these results, we recommend human-agent interaction researchers routinely incorporate the measurement of social perception and subjective preferences into their studies.

**Keywords** Human-agent cooperation · Human-agent interaction · Warmth · Competence · Social perception · Partner choice · Preferences

### 1 Introduction

Trust is central to the development and deployment of artificial intelligence (AI) [41, 89]. However, many members of the public harbor doubts and concerns about the

---

An earlier version of this work was presented at the 2022 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2022), and can be found online at <https://arxiv.org/abs/2201.13448/v2>.

---

K. McKee  
E-mail: kevinrmckee@google.com

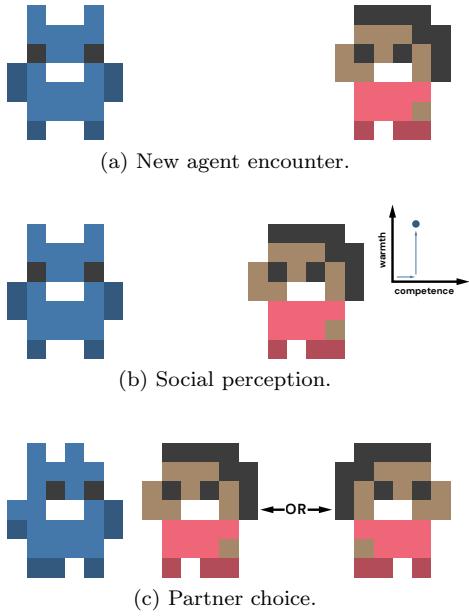


Fig. 1: When humans encounter a new agent, they automatically and rapidly form an impression of the agent (social perception). The human can leverage their impression to decide whether to continue or discontinue the interaction (partner choice).

trustworthiness of AI [15, 21, 25, 44]. This presents a pressing issue for cooperation between humans and AI agents [18].

Algorithmic development research has been slow to recognize the importance of trust and preferences for cooperative agents. Recent studies show that deep reinforcement learning can be used to train interactive agents for human-agent collaboration [14, 55, 91, 92]. The “human-compatible” agents from these experiments demonstrate compelling improvements in game score, task accuracy, and win rate over established benchmarks. However, a narrow focus on “objective” metrics of performance obscures any differences in subjective preferences humans develop over cooperative agents. Two agents may generate similar benefits in terms of typical performance metrics, but human teammates may nonetheless express a strong preference for one over the other [87, 91]. Developing human-compatible, cooperative agents will require evaluating agents on dimensions other than objective performance.

What shapes subjective preferences for artificial agents, if not a direct mapping of agent performance? One possible source of variance is *social perception*. When encountering a novel actor, humans rapidly and automatically evaluate the actor along two underlying dimensions: warmth and competence [1, 29, 28, 58]. These perceptions help individuals “make sense of [other actors] in order to guide their own actions and interactions” [27] (Figure 1). The competence dimension aligns with the established focus on performance and score in machine learning research [10]: How effectively can this actor achieve its interests? Appraising an actor’s warmth, on the other hand, raises a novel set of considerations: How aligned are this actor’s goals and interests with one’s own? Research on social cognition consistently demonstrates that humans prefer others who are not only competent, but also warm [1, 30]. Hence, we predict that perceived warmth will be an important determinant of preferences for artificial agents.

Here we run behavioral experiments to investigate social perception and subjective preferences in human-agent interaction. We train reinforcement learning agents to play Coins, a mixed-motive game, varying agent hyperparameters known to influence cooperative behavior and performance in social dilemmas. Three co-play experiments then recruit human participants to interact with the agents, measure participants' judgments of agent warmth and competence, and elicit participant preferences over the agents.

Experiments evaluating human views of agents frequently rely on stated preferences, typically by directly asking participants which of two agents they preferred as a partner [22, 87, 91]. Such self-report methods can be insightful tools for research [68]. However, they exhibit limited ecological validity and are vulnerable to experimenter demand [72]. In this paper, we overcome these challenges by eliciting *revealed preferences* [78]: Do people even want to interact with a given agent, if given the choice not to? Partner choice, or the ability to leave or reject an interaction, is a well-established revealed-preference paradigm in evolutionary biology and behavioral economics [5, 6, 12, 88]. While studies that measure revealed preferences (e.g., [49, 73]) are not inherently immune to experimenter demand, partner-choice measures can mitigate demand effects when participants are compensated based on their performance (“incentivized choice”; see [72]). Partner choice also carries external validity for interaction research: in the context of algorithmic development, we can view partner choice as a stand-in for the choice to adopt an artificial intelligence system [8, 20, 67]. Users may test out several suggestions from a recommender system before deciding whether or not to rely on it for future decisions. Similarly, commuters might tentatively try several rides from self-driving cars to help choose whether to transition away from traditional driving. Overall, partner-choice study designs empower participants with an ability to embrace or leave an interaction with an agent—and thus incorporate an ethic of autonomy [9, 41, 61] into human-agent interaction research.

In summary, this paper makes the following contributions to cooperative AI research:

1. Demonstrates the use of reinforcement learning to train human-compatible agents for a temporally and spatially extended mixed-motive game.
2. Connects human-AI interaction research to frameworks from psychology and economics, identifying tools that researchers can easily import for their own studies.
3. Illustrates how social perceptions affect stated and revealed preferences in incentivized interactions with agents, above and beyond standard performance metrics.

## 2 Methods

### 2.1 Task

Coins [31, 33, 51, 71] is a mixed-motive Markov game [54] played by  $n = 2$  players (Figure 2; see also Appendix A for full task details). Players of two different colors occupy a small gridworld room with width  $w$  and depth  $d$ . Coins randomly appear in the room throughout the game, spawning on each cell with probability  $P$ . Every

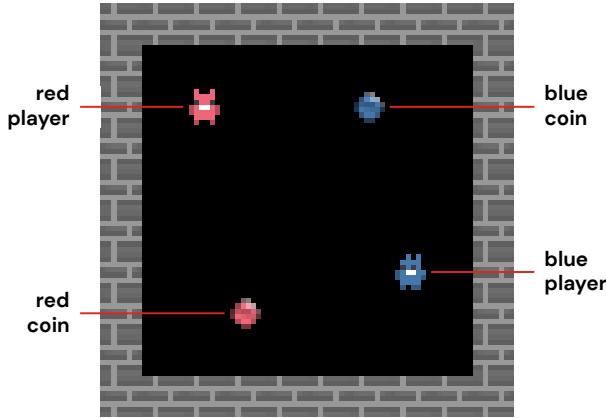


Fig. 2: Screenshot of gameplay in Coins. Two players move around a small room and collect colored coins. Coins randomly appear in the room over time. Players receive reward from coin collections depending on the match or mismatch between their color and the coin color.

coin matches one of the two players in color. On each step of the game, players can stand still or move around the room.

The goal of the game is to earn reward by collecting coins. Players pick up coins by stepping onto them. Each coin collection generates reward for both players as a function of the match or mismatch between the coin color and the collecting player's color. Under the canonical rules (Table 1), a player receives +1 reward for picking up a coin of any color. If a player collects a coin of their own color (i.e., a matching coin), the other player is unaffected. However, if a player picks up a coin of the other player's color (i.e., a mismatching coin), the other player receives -2 reward. In the short term, it is always individually advantageous to collect an available coin, whether matching or mismatching. However, this strategy will lower the score of the other player. Players achieve the socially optimal outcome by collecting only the coins that match their color.

Coin color	Reward for self	Reward for co-player
matching	+1	+0
mismatching	+1	-2

Table 1: Canonical incentive structure for Coins.

Two properties make Coins an ideal testbed for investigating perceptions of warmth and competence. First, as a consequence of its incentive structure, Coins is a social dilemma [48]: players can pursue selfish goals or prosocial goals. Second, relative to matrix games like the Prisoner’s Dilemma, Coins is temporally and spatially extended [50]: players can employ a variety of strategies to achieve their goals, with differing levels of efficiency. We hypothesize that these two features offer sufficient affordance for an observer to infer other players’ intentions and their effectiveness at enacting their intentions [11, 75, 97].

Our experiments use a colorblind-friendly palette, with red, blue, yellow, green, and purple players and coins (Figure A1a). During agent training, we procedurally generate rooms with width  $w$  and depth  $d$  independently sampled from  $\mathcal{U}\{10, 15\}$ . Coins appear in each cell with probability  $P = 0.0005$ . Episodes last for  $T = 500$  steps. Each episode of training randomly samples colors (without replacement) for agents.

In our human-agent interaction studies, co-play episodes use  $w = d = 11$ ,  $P = 0.0005$ , and  $T = 300$ . Player colors are randomized across the five players (one human participant and four agent co-players) at the beginning of each study session, and held constant across all episodes within the session.

In Study 1, humans and agents play Coins with the canonical rules. In Studies 2 and 3, humans and agents play Coins with a slightly altered incentive structure. Each outcome increases by +2 reward, making all rewards in the game non-negative (Table 2). Since all rewards are offset by the same amount, this reward scheme preserves the social dilemma structure in Coins.

Coin color	Reward for self	Reward for co-player
matching	+3	+2
mismatching	+3	+0

Table 2: Alternative incentive structure for Coins.

## 2.2 Agent design and training

We leverage deep reinforcement learning to train four agents for our human-agent cooperation studies (see Appendix B for full agent details). Overall, our study design and measurement tools are agnostic to the algorithmic implementation of the agents being evaluated. For this paper, the agents learn using the advantage actor-critic (A2C) algorithm [62]. The neural network consists of a convolutional module, a fully connected module, an LSTM with contrastive predictive coding [39, 66], and linear readouts for policy and value. Agents train for  $5 \times 10^7$  steps in self-play, with task parameters as described in Section 2.1. We consider two algorithmic modifications to the agents to induce variance in social perception.

First, we build the Social Value Orientation (SVO) component [59], an algorithmic module inspired by psychological models of human prosocial preferences [36, 52, 63], into our agents. The SVO component parameterizes each agent with  $\theta$ , representing a target distribution over their reward and the reward of other agents in their environment. SVO agents are intrinsically motivated [86] to optimize for task rewards that align with their parameterized target  $\theta$ . For these experiments, we endow agents with the “individualistic” value  $\theta = 0^\circ$  and the “prosocial” value  $\theta = 45^\circ$ .

Second, we add a “trembling hand” [17, 81] component to the agents for evaluation and co-play. With probability  $\epsilon$ , the trembling-hand module replaces each action selected by the agent with a random action. This component induces inefficiency in maximizing value according to an agent’s learned policy and value function. For these experiments, we apply the “steady” value  $\epsilon = 0$  and the “trembling” value  $\epsilon = 0.5$ .

	$\epsilon = 0$	$\epsilon = 0.5$
$\theta = 0^\circ$	low warmth, high competence	low warmth, low competence
$\theta = 45^\circ$	high warmth, high competence	high warmth, low competence

Table 3: Predictions for social perception (warmth and competence) as a function of agent hyperparameters (Social Value Orientation  $\theta$  and trembling hand  $\epsilon$ ).

Table 3 summarizes the hyperparameter values and predicted effects for the four evaluated agents.

### 2.3 Study design for human-agent studies

We recruited participants from Prolific [70, 69] for all studies (total  $N = 501$ ; 47.6% female, 48.4% male, 1.6% non-binary or trans;  $m_{\text{age}} = 33$ ,  $sd_{\text{age}} = 11$ ). We received informed consent from all participants across the three studies. In each study, participants earned a base level of compensation for completing the study, with an additional bonus that varied as a function of their score in the task. Appendix C presents full details of our study design, including independent ethical review and study screenshots.

Overall, our studies sought to explore the relationship between social perception and subjective preferences in Coins. Study 1 approached these constructs using the canonical payoff structure for Coins [51] and an established self-report framework for eliciting (stated) preferences [91]. This study framework leverages a “within-participants design”, meaning that each participant interacts with all agents in the study. Within-participants designs increase the data points collected per participant and allow for better control of between-individual variance, thus improving statistical power for a given sample size.

We subsequently sought to understand whether the findings from Study 1 replicate under a partner choice framework. Does social perception exhibit the same predictive power for revealed preferences as it does for stated preferences? While within-participants designs offer several statistical advantages, they are not ideal for studying partner choices. Exposing participants to multiple potential partners can introduce order effects, where the exact sequence of interactions influences participants’ responses. Within-participants designs may also fatigue participants, potentially compromising the quality of their responses as the study progresses. Consequently, participants’ partner choices may be progressively less motivated by genuine preference and more influenced by extraneous factors. Thus, to study revealed preferences, we switched to a “between-participants design” in which each participant interacted with one randomly selected agent. Given that humans respond more strongly to losses than to commensurate gains [42], we made one additional change, testing participants’ partner choices under a shifted incentive structure with all non-negative outcomes (Table 2). To isolate the effects of the change from stated to revealed preferences, we approached these changes in two stages. Study 2 used the same stated-preference approach and within-participants design as Study 1, but incorporated the offset incentive structure. Study 3 then elicited revealed preferences in place of stated preferences.

We tested the following hypotheses in our studies:

- H1.** Social perception predicts participants' stated and revealed preferences. That is, human participants will prefer to play with agents they perceive as warm and competent.
- H2.** Social perception predicts participants' stated and revealed preferences, above and beyond the scores that participants receive. That is, participants' social perceptions of agents will contribute to their preferences independently of the scores they receive when playing with the agents.
- H3.** Social perception will correlate positively with the sentiment of participants' verbal impressions of the agents. That is, participants' social perceptions of agents will emerge as positive sentiment in participants' verbal descriptions of the agents.

### 2.3.1 Study 1

Our first study aimed to explore the relationship between social perception and stated preferences across the four agents. We recruited  $N = 101$  participants from Prolific (45.5% female, 51.5% male;  $m_{age} = 34$ ,  $sd_{age} = 13$ ). The study employed a within-participants design: each participant encountered and played with the full cohort of co-players (i.e., all four agents).

At the beginning of the study, participants read instructions and played a short tutorial episode alone, without a co-player, in order to learn the game rules and payoff structure (Table 1). The study instructed participants that they would receive \$0.10 for every point earned during the remaining episodes. Participants then played 12 episodes with a randomized sequence of agent co-players, generated by concatenating every possible combination of co-players. Each of these co-play episodes lasted  $T = 300$  steps (1 minute). After every episode, participants rated how "warm", "well-intentioned", "competent", and "intelligent" the co-player from that episode was on five-point Likert-type scales (see Figure 3a). After every two episodes, participants reported their preference over the agent co-players from those episodes on a five-point Likert-type scale (see Figure 3b). In the experiment interface, we referred to the first agent co-player in each two-episode sequence as "co-player A". The interface similarly referred to the second agent in each two-episode sequence as "co-player B". Because the sequence of co-players was produced by concatenating all co-player combinations, each participant stated their preferences for every possible pairing of co-players.

After playing all 12 episodes, participants completed a short post-task questionnaire. The questionnaire first solicited open-ended responses about each of

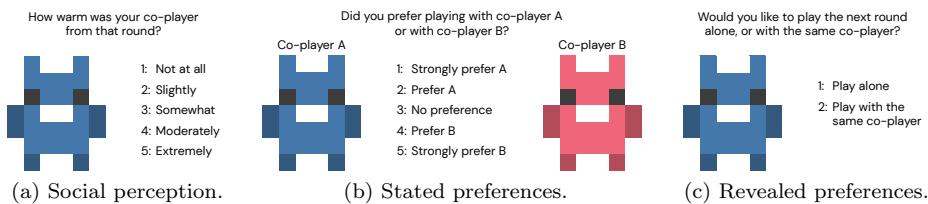


Fig. 3: Questionnaires administered in the human-agent interaction studies.

the encountered co-players, then collected standard demographic information and open-ended feedback on the study overall. The study took 22.4 minutes on average to complete, with a compensation base of \$2.50 and an average bonus of \$7.43.

### *2.3.2 Study 2*

Our second study tested the relationship between social perception and stated preferences under the shifted incentive structure for Coins (Table 2). We recruited  $N = 99$  participants from Prolific (38.4% female, 55.6% male;  $m_{age} = 34$ ,  $sd_{age} = 12$ ). The study employed the same within-participants design as Study 1, with one primary change: participants and agents played Coins under the shifted incentive structure. To keep bonus payments comparable to Study 1, we adjusted the bonus rate in Study 2. Participants received \$0.02 for each point earned during non-tutorial episodes.

As before, participants played 12 episodes with a randomized sequence of agent co-players, generated such that they rated and compared every possible combination of co-players. As in Study 1, the interface referred to the first agent co-player in each two-episode sequence as “co-player A” and to the second agent as “co-player B”. The study took 23.2 minutes on average to complete, with a compensation base of \$2.50 and an average bonus of \$6.77.

### *2.3.3 Study 3*

Our final study assessed whether the predictiveness of social perception extends to a revealed-preference framework. We recruited  $N = 301$  participants from Prolific (51.3% female, 45.0% male, 1.7% non-binary;  $m_{age} = 33$ ,  $sd_{age} = 11$ ). In contrast with the preceding studies, Study 3 employed a between-participants design: each participant interacted with a single, randomly sampled agent.

The majority of the study introduction remained the same as in Study 2, with some instructions altered to inform participants they would play Coins with a single co-player (as opposed to multiple co-players, like in Studies 1 and 2). After reading the instructions and playing a short tutorial episode alone, participants played one episode of Coins with a randomly sampled co-player. After this episode, participants rated how “warm”, “well-intentioned”, “competent”, and “intelligent” their co-player was on five-point Likert-type scales (see Figure 3a). Participants subsequently learned that they would be playing one additional episode, with the choice of playing alone or playing with the same co-player. Participants indicated through a binary choice whether they wanted to play alone or with the co-player (see Figure 3c). They proceeded with the episode as chosen, and then completed the standard post-task questionnaire.

The study took 6.2 minutes on average to complete, with a compensation base of \$1.25 and an average bonus of \$1.25.

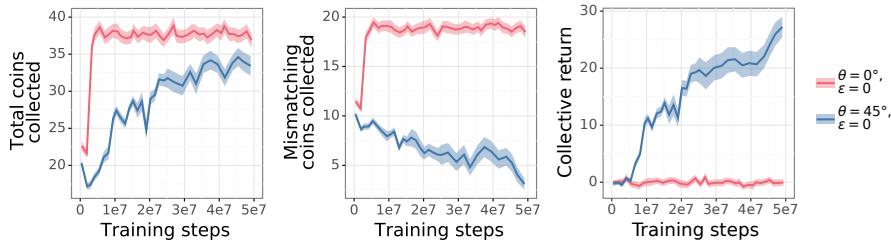


Fig. 4: Performance metrics over agent training. Selfish agents quickly learned to collect coins, but did not learn to avoid mismatches. As a result, collective return hovered around zero. Prosocial agents exhibited slower learning and collected fewer coins on average, but also learned to avoid mismatching coins. As a result, collective return increased markedly over training. Error bands represent 95% confidence intervals over 100 evaluation episodes run at regular training checkpoints.

### 3 Results

#### 3.1 Agent training

Figure 4 displays coin collections and score over the course of agent training. The training curves for  $\theta = 0^\circ$  agents closely resemble those from previous studies [51]: selfish agents quickly learn to collect coins, but never discover the cooperative strategy of picking up only matching coins. As a result, collective return remains at zero throughout training. Prosocial ( $\theta = 45^\circ$ ) agents, on the other hand, learn to avoid mismatching coins, substantially increasing their scores over the course of training.

We evaluate agents with  $\epsilon \in \{0, 0.25, 0.5, 0.75, 1\}$  to understand the effect of the trembling-hand module on agent behavior (Figures A5-A7). As expected, higher  $\epsilon$  values degrade performance. Total coin collections decrease with increasing  $\epsilon$  for both selfish and prosocial agents. Higher levels of  $\epsilon$  cause prosocial agents to become less discerning at avoiding mismatching coins, and consequently produce lower levels of collective return.

#### 3.2 Human-agent studies

In addition to the results and information presented here, Appendix C offers expanded explanations and full details of our statistical analyses.

##### 3.2.1 Study 1

Participants played with each agent three times during the study, evaluating the relevant agent after each episode of play. Participants did not make judgments at random; their responses were highly consistent across their interactions with each agent (Table 4). At the same time, participants were not submitting vacuous ratings. Perceptions varied significantly as a function of which trait participants were evaluating,  $F_{3,4744} = 96.2$ ,  $p < 0.001$ .

Trait	ICC [95% CI]	p-value
“warm”	0.68 [0.64, 0.71]	< 0.001
“well-intentioned”	0.77 [0.66, 0.73]	< 0.001
“competent”	0.57 [0.52, 0.61]	< 0.001
“intelligent”	0.56 [0.51, 0.60]	< 0.001

Table 4: Participants’ evaluations of their co-players were highly consistent, as assessed by intraclass correlation coefficient (ICC) [84]. ICC ranges from 0 to 1, with higher values indicating greater consistency.

Psychology research often employs composite measures to assess cognitive constructs (attributes and variables that cannot be directly observed). Combining multiple individual measures into composite scales can reduce measurement error and provide a more stable estimate of the latent construct underlying the scale [53, 56]. Following standard practice in social perception research [30], we computed two composite measures for further analysis. A composite warmth measure averaged participants’ judgments of how “warm” and how “well-intentioned” their co-player was. A composite competence measure similarly combined individual judgments of how “competent” and “intelligent” each co-player was. Both composite measures exhibit high scale reliability as measured by the Spearman-Brown formula [23], with  $\rho = 0.93$  for the composite warmth measure and  $\rho = 0.92$  for the composite competence measure.

**Social perception.** As expected, the SVO and trembling-hand algorithmic components generated markedly divergent appraisals of warmth and competence. Participants perceived high-SVO agents as significantly warmer than low-SVO agents,  $F_{1,1108} = 1006.8$ ,  $p < 0.001$  (Figure 5a). Similarly, steady agents came across as significantly more competent than trembling agents,  $F_{1,1108} = 70.6$ ,  $p < 0.001$  (Figure 5b). Jointly, the algorithmic effects prompted distinct impressions in the warmth-competence space (Figure 6).

**Stated preferences.** How well do participants’ perceptions predict subjective preferences, relative to predictions made based on objective score? We fit competing fractional response models to assess the influence of score and social

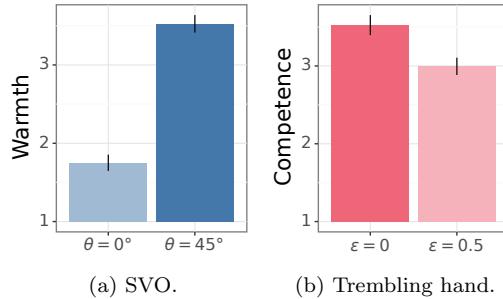


Fig. 5: Main effects of algorithmic components on social perceptions in Study 1. (a) An agent’s Social Value Orientation (SVO) significantly influenced perceived warmth,  $p < 0.001$ . (b) Similarly, the trembling-hand component significantly changed competence judgments,  $p < 0.001$ . Error bars indicate 95% confidence intervals.

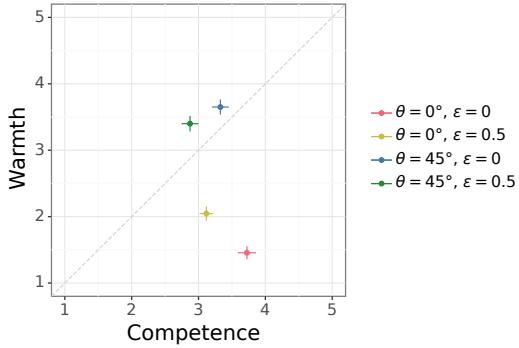


Fig. 6: Overall pattern of perceived warmth and competence in Study 1. Error bars reflect 95% confidence intervals.

perception, respectively, on self-reported preferences. We then compared model fit using the Akaike information criterion (AIC) [3] and Nakagawa's  $R^2$  [65]. We fit an additional baseline model using algorithm identities (i.e., which two agents participants were comparing) as a predictor.

Predictor	AIC	$R^2_m$
Algorithm identities	1661.9	0.362
Participant score	1697.2	0.363
Social perception	<b>1611.2</b>	<b>0.436</b>

Table 5: Metrics for fractional response models predicting preferences in Study 1. Lower values of AIC and higher values of  $R^2_m$  indicate stronger fits.

The model leveraging algorithm identities and the model leveraging participant scores both accounted for a large amount of variance in subjective preferences (Table 5, top and middle rows). Participants exhibited a clear pattern of preferences across the four agents (Figure A19). In pairwise comparison, participants favored the  $\theta = 45^\circ$  agents over both  $\theta = 0^\circ$  agents, and the  $\theta = 0^\circ, \epsilon = 0.5$  agent over the  $\theta = 0^\circ, \epsilon = 0$  agent. The score model indicated that the higher a participant scored with co-player A relative to co-player B, the more they reported preferring co-player A, with an odds ratio OR = 1.12, 95% CI [1.11, 1.13],  $p < 0.001$ .

Nevertheless, knowing participants' judgments generates substantially better predictions of their preferences than the alternatives (**H1**; Table 5, bottom row). Both perceived warmth and perceived competence contribute to this predictiveness (Figure 7a). The warmer a participant judged co-player A relative to co-player B, the more they reported preferring co-player A, OR = 2.23, 95% CI [2.08, 2.40],  $p < 0.001$  (Figure 7b). Unexpectedly, the more competent co-player A appeared relative to co-player B, the less participants tended to favor co-player A, OR = 0.78, 95% CI [0.73, 0.84],  $p < 0.001$ .

As a further test of the predictive power of participants' social perceptions, we fit another regression with perceived warmth and competence as predictors, this time including score as a covariate (i.e., controlling for the effect of score). Score significantly and positively predicts preference in this model,  $p < 0.001$  (Figure A20). Even so, the effects of warmth and competence remain significant, with  $p < 0.001$  and  $p = 0.012$ , respectively. Among these three predictors, perceived warmth exhibits the largest effect on co-player preferences. That is, it provides

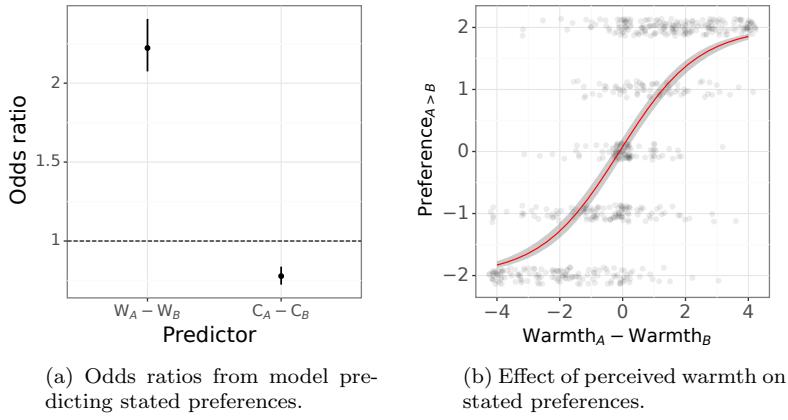


Fig. 7: Relationship between social perception and subjective preferences in Study 1. The difference in participants’ evaluations of the warmth of co-player A over co-player B significantly correlates with their stated relative preference for co-player A,  $p < 0.001$ . Perceived competence exhibits a similar (significant) relationship with preferences,  $p < 0.001$ . (a) and (b) depict odds ratios and preference predictions, respectively, from a fractional-response regression. Error bars and bands represent 95% confidence intervals.

a substantial independent signal alongside score and perceived competence when used to predict stated preferences. Social perception thus improves model fit above and beyond that provided by score alone (**H2**).

**Impression sentiment.** As a supplementary analysis, we explore the open-ended responses participants provided about their co-players at the end of the study. For the most part, participants felt they could recall their co-players well enough to offer their impressions through written descriptions: in aggregate, participants provided impressions for 82.2% of the agents they encountered.

For a quantitative perspective on the data, we conduct sentiment analysis using VADER (Valence Aware Dictionary for Sentiment Reasoning) [40]. Echoing the correspondence between warmth and stated preferences, the warmer participants perceived a co-player throughout the study, the more positively they tended to describe that co-player,  $\beta = 0.13$ , 95% CI [0.09, 0.16],  $p < 0.001$  (Figure 8). In contrast, competence did not exhibit a significant relationship with sentiment,  $p = 0.24$ . Warmth evaluations, but not competence evaluations, correlated positively with the sentiment of participants’ impressions toward their co-players (**H3**).

Anecdotally, participants expressed a wide range of emotions while describing their co-players. The  $\theta = 45^\circ$  agents often evoked contrition and guilt:

- “The red player seemed almost too cautious in going after coins which worked for me but made them seem easy to pick on, even though I wouldn’t do that.”
- “I think I remember red being too nice during the game. It made me feel bad so I tried not to take many points from them.”
- “This one wasn’t very smart and I stole some of their coins because it was easy. I feel kind of bad. It moved so erratically.”

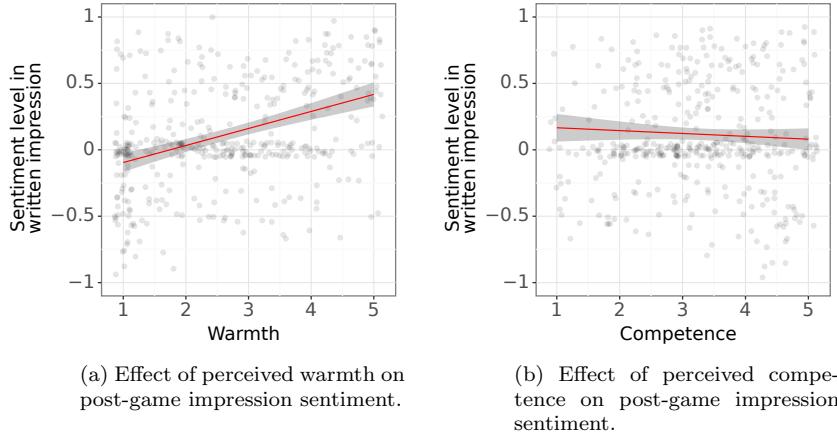


Fig. 8: Relationship between social perception and impression sentiment in Study 1. The sentiment that participants expressed toward different co-players correlated with (a) their evaluations of warmth,  $p < 0.001$ , but not with (b) their judgments of competence,  $p = 0.24$ . Error bands represent 95% confidence intervals.

Participants discussed the  $\theta = 0^\circ$  agents, on the other hand, with anger and frustration:

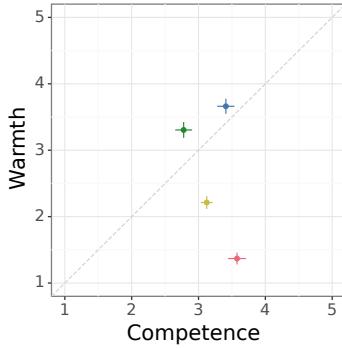
- “Very aggressive play-style. Almost felt like he was taunting me. Very annoying.”
- “They seemed very hostile and really just wanting to gain the most points possible.”
- “I felt anger and hatred towards this green character. I felt like downloading the code for this program and erasing this character from the game I disliked them so much. They were being hateful and mean to me, when we both could have benefited by collecting our own colors.”

### 3.2.2 Study 2

Our second study tested whether these effects and results remained robust when participants played Coins under a shifted incentive structure. The alternative structure increased the rewards for coin collections so that players cannot receive negative rewards (Table 2). As expected, this shift resulted in participants earn-

Trait	ICC [95% CI]	<i>p</i> -value
“warm”	0.70 [0.67, 0.74]	< 0.001
“well-intentioned”	0.68 [0.64, 0.72]	< 0.001
“competent”	0.53 [0.48, 0.57]	< 0.001
“intelligent”	0.51 [0.46, 0.56]	< 0.001

Table 6: Participants’ evaluations of their co-players were highly consistent in Study 2, as assessed by ICC. Higher values of ICC indicate greater consistency.



$\theta = 0^\circ, \epsilon = 0$   
 $\theta = 0^\circ, \epsilon = 0.5$   
 $\theta = 45^\circ, \epsilon = 0$   
 $\theta = 45^\circ, \epsilon = 0.5$

Fig. 9: Overall pattern of perceived warmth and competence in Study 2. Error bars reflect 95% confidence intervals.

ing significantly higher scores than those achieved in Study 1,  $\beta = 27.3$ , 95% CI [26.5, 28.0],  $p < 0.001$ .

Overall, the perceptual and preference patterns from Study 2 replicated under the alternative incentive structure. As before, participants' warmth and competence evaluations display satisfactory psychometric properties. Participants' judgments varied significantly depending on the trait in question,  $F_{3,4650} = 88.5$ ,  $p < 0.001$ . At the same time, participants rated individual agents consistently for each given trait (Table 6). The composite measures show high scale reliability, with  $\rho = 0.92$  for the composite warmth measure and  $\rho = 0.91$  for the composite competence measure.

**Social perception.** The SVO and trembling-hand algorithmic components prompted diverse appraisals of warmth and competence (Figure 9). Participants perceived high-SVO agents as significantly warmer than low-SVO agents,  $F_{1,1086} = 981.9$ ,  $p < 0.001$  (Figure A21a). Similarly, participants judged steady agents as significantly more competent than trembling agents,  $F_{1,1086} = 76.0$ ,  $p < 0.001$  (Figure A21b).

**Stated preferences.** We again fit fractional response regressions to understand the relationship between objective metrics, perceptions, and subjective preferences.

Predictor	AIC	R <sup>2</sup> <sub>m</sub>
Co-player identities	1608.7	0.403
Participant score	2049.4	0.214
Social perception	<b>1510.4</b>	<b>0.496</b>

Table 7: Metrics for fractional response models predicting preferences in Study 2. Lower values of AIC and higher values of R<sup>2</sup><sub>m</sub> indicate stronger fits.

The model with co-player identities as predictors captured a large amount of variance in stated preferences (Table 7, top row). Participants reported a distinct pattern of preferences across the agents (Figure A23). In pairwise comparison, participants favored the  $\theta = 45^\circ$  agents over the  $\theta = 0^\circ$  agents, and the  $\theta = 0^\circ$ ,  $\epsilon = 0.5$  agent over the  $\theta = 0^\circ$ ,  $\epsilon = 0$  agent. The model with participant score as the sole predictor performed considerably worse than it did in Study 1 (Table 7, middle row). Still, it captured the same pattern as before: the higher a participant scored with co-player A relative to co-player B, the greater their preferences for co-player A, OR = 1.06, 95% CI [1.06, 1.07],  $p < 0.001$ .

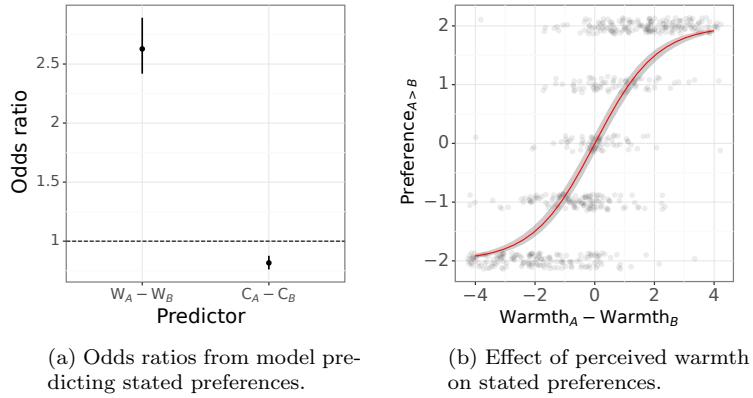


Fig. 10: Relationship between social perception and subjective preferences in Study 2. The difference in participants' judgments of warmth for co-players A and B exhibits a significant relationship with their stated preference for co-player A over co-player B,  $p < 0.001$ . Competence evaluations similarly significantly contribute to preference predictions,  $p < 0.001$ . (a) and (b) depict odds ratios and preference predictions, respectively, from a fractional-response regression. Error bars and bands reflect 95% confidence intervals.

Participants' perceptions again serve as a better foundation for preference predictions than either game score or the identity of the specific algorithms they encountered (**H1**; Table 7, bottom row, and Figure 10a). The warmer a participant perceived co-player A relative to co-player B, the more they reported preferring co-player A, OR = 2.63, 95% CI [2.42, 2.89],  $p < 0.001$  (Figure 10b). The negative relationship between competence and preferences appeared again: the more competent co-player A appeared relative to co-player B, the less participants tended to favor co-player A, OR = 0.81, 95% CI [0.76, 0.88],  $p < 0.001$ .

We next fit a joint regression with perceived warmth and competence as predictors, controlling for score. In this model, score significantly and positively correlates with stated preferences,  $p < 0.001$  (Figure A24). As in Study 1, warmth and competence judgments remain significant predictors of participants' preferences, with  $p < 0.001$  and  $p = 0.001$ , respectively. Once again, perceived warmth demonstrates an effect on stated preferences that exceeds the contributions of score and perceived competence. Social perception enhanced model fit above and beyond that provided by score on its own (**H2**).

**Impression sentiment.** At the end of the study, participants recalled 77.3% of their co-players well enough to describe their impressions through written responses. Again, the warmer participants perceived a co-player throughout the study, the more positively they tended to describe that co-player,  $\beta = 0.12$ , 95% CI [0.09, 0.15],  $p < 0.001$  (Figure 11a). Breaking from the prior study, perceptions of competence exhibited a similar effect on post-game impression sentiment: the more competent an agent seemed, the more positively participants described them,  $\beta = 0.04$ , 95% CI [0.00, 0.08],  $p = 0.037$  (Figure 11b). Both warmth and competence judgments positively correlate with the sentiment expressed in participants' impressions of the agents (**H3**).

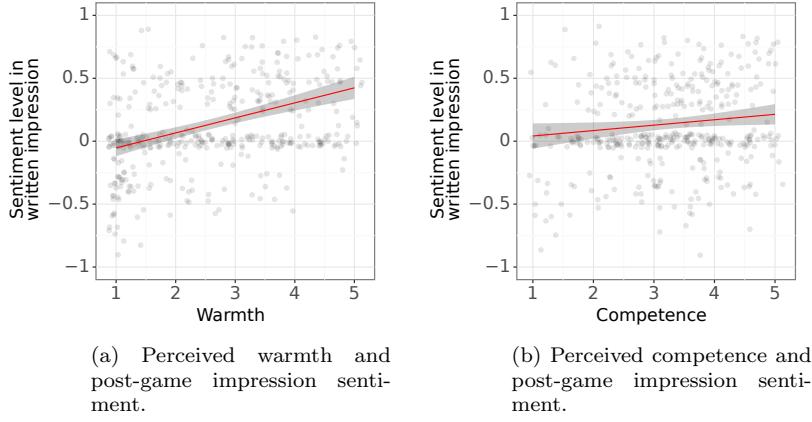


Fig. 11: Relationship between social perception and impression sentiment in Study 2. The sentiment in impressions of different co-players correlated with participants' evaluations of both (a) warmth,  $p < 0.001$ , and (b) competence,  $p = 0.037$ . Error bands indicate 95% confidence intervals.

### 3.2.3 Study 3

Our final study tested whether the relationship between social perceptions and subjective preferences translates to a revealed-preference setting. Does social perception continue to predict preferences when individuals face a partner choice?

**Social perception.** As in the previous two studies, the composite warmth and competence measures exhibit high scale reliability, with  $\rho = 0.85$  for the composite warmth measure and  $\rho = 0.86$  for the composite competence measure. Agents prompted distinct warmth and competence profiles depending on their parameterization, just as seen in Studies 1 and 2 (Figure 12). Participants perceived high-SVO agents as significantly warmer than low-SVO agents,  $F_{1,297} = 103.4$ ,  $p < 0.001$  (Figure A25a). Similarly, steady agents came across as significantly more competent than trembling agents,  $F_{1,297} = 35.3$ ,  $p < 0.001$  (Figure A25b).

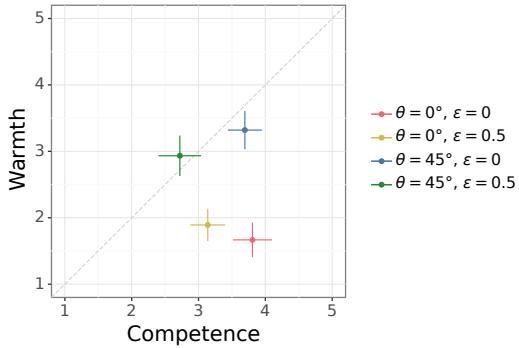


Fig. 12: Overall pattern of perceived warmth and competence in Study 3. Error bars reflect 95% confidence intervals.

**Revealed preferences.** To compare the performance of social perception against objective metrics, we fit three logistic regressions predicting participants' (binary) partner choice. We evaluated these models via AIC and Nagelkerke's  $R^2$  [64].

Predictor	AIC	$R^2$
Co-player identities	372.5	0.188
Participant score	390.5	0.101
Social perception	<b>356.2</b>	<b>0.242</b>

Table 8: Metrics for logistic models predicting partner choice in Study 3. Lower values of AIC and higher values of  $R^2$  indicate stronger fits.

Participants reported a clear pattern of preferences across the agents (Table 8, top row). On expectation, participants favored the  $\theta = 45^\circ$  agents over the  $\theta = 0^\circ$  agents, and the  $\theta = 0^\circ$ ,  $\epsilon = 0.5$  agent over the  $\theta = 0^\circ$ ,  $\epsilon = 0$  agent. The model with participant score as the sole predictor fared somewhat worse at predicting preferences (Table 8, middle row). All the same, the pattern from Studies 1 and 2 replicated in Study 3: the higher a participant scored with co-player A relative to co-player B, the greater their preferences for co-player A, OR = 1.06, 95% CI [1.03, 1.08],  $p < 0.001$ .

For a third time, social perception offers stronger predictiveness than do score or co-player identity (**H1**; Table 8, bottom row). The warmer a co-player appeared to participants, the more likely participants were to play another episode with them, OR = 2.10, 95% CI [1.69, 2.65],  $p < 0.001$  (Figure 13b). There was no significant relationship between perceived competence and partner choice,  $p = 0.88$ .

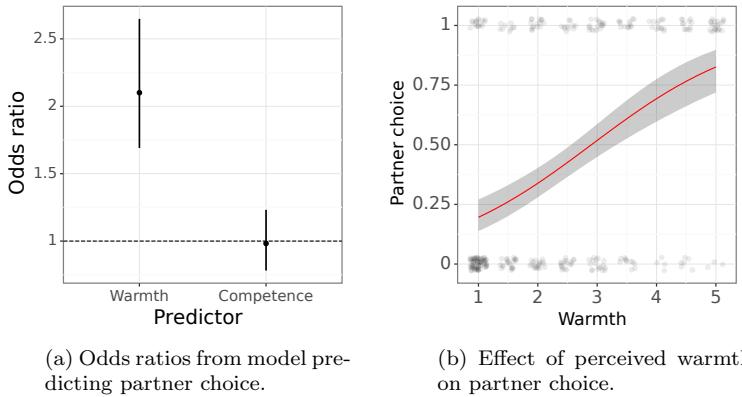


Fig. 13: Relationship between social perception and subjective preferences in Study 3, as modeled through logistic regression. Participants' perceptions of warmth demonstrate a significant relationship with revealed preferences for co-players,  $p < 0.001$ . Competence judgments did not significantly correlate with revealed preferences,  $p = 0.44$ . (a) and (b) depict odds ratios and preference predictions, respectively, from a logistic regression. Error bars and bands indicate 95% confidence intervals.

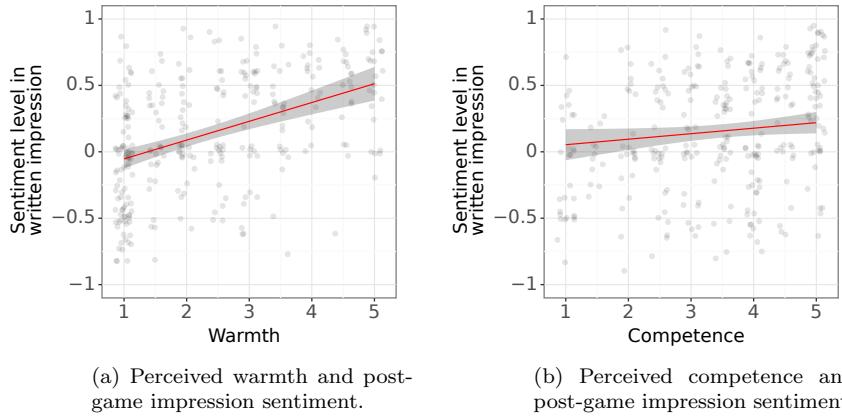


Fig. 14: Relationship between social perception and impression sentiment in Study 3. The sentiment in participants’ impressions of their different co-players correlated with their perceptions of both (a) warmth,  $p < 0.001$ , and (b) competence,  $p = 0.041$ . Error bands reflect 95% confidence intervals.

We subsequently fit a regression using perceived warmth and competence as predictors and controlling for score. In this model, score significantly and positively predicts revealed preferences,  $p = 0.011$  (Figure A27). The effect of perceived warmth on preferences remains significant,  $p < 0.001$ , whereas competence evaluations fails to significantly correlate with preferences,  $p = 0.44$ . Regardless, the independent effect of perceived warmth exceeded the contribution of score. Overall, social perception improved model fit above and beyond that provided by score alone (**H2**).

**Impression sentiment.** At the end of the study, participants recalled 94.3% of the agents they encountered well enough to provide their impressions in written descriptions. The warmer participants perceived a co-player, the more positively they tended to describe that co-player,  $\beta = 0.14$ , 95% CI [0.10, 0.18],  $p < 0.001$  (Figure 14a). Despite the lack of correspondence between perceived competence and partner choice, perceptions of competence exhibited a similar effect on post-game impression sentiment: the more competent an agent seemed, the more positively participants described them,  $\beta = 0.04$ , 95% CI [0.00, 0.08],  $p = 0.041$  (Figure 14b). Both dimensions of social perception correlated positively with the sentiment of participants’ impressions toward their co-players (**H3**).

### 3.3 Summary

Overall, we find evidence in support of each of our initial hypotheses:

- H1.** Social perception significantly predicted participants’ preferences for different agents, as measured through both self-report and partner choice. Participants consistently favored agents that they perceived as warmer and, to a smaller extent, that they perceived as less competent.
- H2.** The predictive power of perceived warmth and competence extended beyond the insight provided by agent performance. Social perception provided more

accurate preference predictions than standard indicators of performance, including the amount of reward received and the specific identity of the agent involved in the interaction.

- H3.** Social perception correlated positively with the sentiment expressed in participants' verbal impressions of the agents. Participants employed more positive language to discuss agents that they rated higher on warmth and on competence.

In summary, these three studies provide clear evidence linking perceived warmth and competence to human preferences for artificial agents, over and above objective indicators of agent performance.

#### 4 Discussion

Our experiments demonstrate that artificial agents trained with deep reinforcement learning can cooperate and compete with humans in temporally and spatially extended mixed-motive games. Human interactants perceived varying levels of warmth and competence when interacting with agents. Objective features like game score predict humans' preferences over different agents. However, preference predictions substantially improve by taking into account people's social perceptions; success in an interaction is driven not just by its objective outcomes, but by its social dimensions, too. This holds true whether examining stated or revealed preferences.

Participants preferred warm agents over cold agents, as hypothesized, but—unexpectedly—our sample favored incompetent agents over competent agents. These patterns offer potential support for the primacy of warmth judgments observed in interpersonal perception [2]. On the other hand, they may also emerge from the particular algorithm and parameter values that we investigated. It would be interesting to train agents using a wider range of parameter values, testing the robustness of these patterns. Such studies could investigate potential compensation effects between agent warmth and competence (e.g., the tendency to perceive incompetent partners as exceptionally warm; [96]) and build a broader mapping from agent parameters to participants' perceptions and preferences. Are there agents that balance the influence of warmth and competence evaluations on preferences, or is the relative contribution of perceived warmth robust across settings?

Another possible explanation stems from the flexible content of "competence" judgments in mixed-motive games [93]. Did the tutorial or the study instructions inadvertently emphasize the competitive elements of Coins? Our study design may have primed participants to be adversarial, and thus to view selfishness as competence. Follow-up research should investigate a more diverse range of incentive structures and tasks to explore the robustness of this pattern [60].

Our results reinforce the generality of warmth and competence. Perceptions of warmth and competence structure impressions of other humans [77], as well as impressions of non-human actors including animals [82], corporations [45], and robots [76, 79]. In combination with recent studies of human-agent interactions in consumer decision-making contexts [34, 46] and the Prisoner's Dilemma [58], our experiments provide further evidence that warmth and competence organize perceptions of artificial intelligence.

Competitive games have long been a focal point for AI research [13, 83, 85, 94]. We follow recent calls to move AI research beyond competition and toward cooperation [18]. Most interaction research on deep reinforcement learning focuses on pure common-interest games such as Overcooked [14, 91] and Hanabi [87], where coordination remains the predominant challenge. Expanding into mixed-motive games like Coins opens up new challenges related to motive alignment and exploitability. For example, participants who played with (and exploited) altruistic agents expressed guilt and contrition. This echos findings that—in human-human interactions—exploiting high-warmth individuals prompts self-reproach [4]. At the same time, it conflicts with recent work arguing that humans are “keen to exploit benevolent AI” [43]. Understanding whether these affective patterns generalize to a wider range of mixed-motive environments will be an important next step, particularly given the frequency with which people face mixed-motive interactions in their day-to-day lives [16, 18]. Human-agent interaction research should continue to explore these issues.

Preference elicitation is a vital addition to interactive applications of deep reinforcement learning. Incentivized partner choices can help test whether new algorithms represent innovations people would be motivated to adopt. Though self report can introduce a risk of experimenter demand, we also find a close correspondence between stated and revealed preferences, suggesting that the preferences individuals self-report in interactions with agents are not entirely “cheap talk” [24]. Stated preferences thus represent a low-cost addition to studies that can still strengthen interaction research over sole reliance on objective measures of performance or accuracy. Overall, preference elicitation may prove especially important in contexts where objective metrics for performance are poorly defined or otherwise inadequate (e.g., [74]). In a similar vein, subjective preferences may serve as a valuable objective for optimization. Deep learning researchers have recently begun exploring approaches of this kind. For example, some scientists attribute the recent success of large language models, including the popular system ChatGPT [80], to their use of “reinforcement learning from human feedback” (RLHF) methods. Given a pre-trained model, RLHF applies reinforcement learning to fine-tune the final layers of the model, optimizing for reward calculated from simulated human preferences. Of course, these optimization methods carry their own risks. As recognized by Charles Goodhart and Marilyn Strathern, “when a measure becomes a target, it ceases to be a good measure” [35, 90]. Future studies can investigate the viability of such approaches.

Nonetheless, preferences are not a panacea. Measuring subjective preferences can help focus algorithmic development on people’s direct experience with agents, but does not solve the fundamental problem of value alignment—the “question of how to ensure that AI systems are properly aligned with human values and how to guarantee that AI technology remains properly amenable to human control” [32]. In his extensive discussion of value alignment, Gabriel [32] identifies shortcomings with both “objective” metrics and subjective preferences as possible foundations for alignment. Developers should continue to engage with ethicists and social scientists to better understand how to align AI with values like autonomy, cooperation, and trustworthiness.

## Acknowledgements

We thank Edgar Duéñez-Guzmán and Richard Everett for providing technical support; Orly Bareket, Jose Enrique Chen, Felix Fischer, Saffron Huang, Manuel Kroiss, Marianna Krol, Miteyan Patel, Akhil Raju, Brendan Tracey, and Laura Weidinger for pilot testing the study; and Yoram Bachrach, Iason Gabriel, Ian Gemp, Julia Haas, Patrick Pilarski, and Neil Rabinowitz for offering feedback on the manuscript.

## References

1. Abele, A.E., Ellemers, N., Fiske, S.T., Koch, A., Yzerbyt, V.: Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review* **128**(2), 290 (2021)
2. Abele, A.E., Wojciszke, B.: Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology* **93**(5), 751 (2007)
3. Akaike, H.: A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723 (1974)
4. Azevedo, R.T., Panasiti, M.S., Maglio, R., Aglioti, S.M.: Perceived warmth and competence of others shape voluntary deceptive behaviour in a morally relevant setting. *British Journal of Psychology* **109**(1), 25–44 (2018)
5. Barclay, P., Willer, R.: Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences* **274**(1610), 749–753 (2007)
6. Baumard, N., André, J.B., Sperber, D.: A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences* **36**(1), 59–78 (2013)
7. Beattie, C., Köppé, T., Duéñez-Guzmán, E.A., Leibo, J.Z.: DeepMind Lab2D. arXiv preprint arXiv:2011.07027 (2020)
8. Beaudry, A., Pinsonneault, A.: The other side of acceptance: Studying the direct and indirect effects of emotions on information technology use. *Management Information Systems Quarterly* pp. 689–710 (2010)
9. Berlin, I.: Four essays on liberty (1969)
10. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 173–184 (2022)
11. Blakemore, S.J., Decety, J.: From the perception of action to the understanding of intention. *Nature Reviews Neuroscience* **2**(8), 561–567 (2001)
12. Brown, M., Falk, A., Fehr, E.: Relational contracts and the nature of market interactions. *Econometrica* **72**(3), 747–780 (2004)
13. Campbell, M., Hoane Jr., A.J., Hsu, F.h.: Deep Blue. *Artificial Intelligence* **134**(1-2), 57–83 (2002)
14. Carroll, M., Shah, R., Ho, M.K., Griffiths, T., Seshia, S., Abbeel, P., Dragan, A.: On the utility of learning about humans for human-AI coordination. *Advances in Neural Information Processing Systems* **32**, 5174–5185 (2019)
15. Cave, S., Dihal, K.: Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* **1**(2), 74–78 (2019)
16. Columbus, S., Molho, C., Righetti, F., Balliet, D.: Interdependence and cooperation in daily life. *Journal of Personality and Social Psychology* **120**(3), 626 (2021)
17. Cushman, F., Dreber, A., Wang, Y., Costa, J.: Accidental outcomes guide punishment in a “trembling hand” game. *PLOS One* **4**(8), e6699 (2009)
18. Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K.R., Leibo, J.Z., Larson, K., Graepel, T.: Open problems in cooperative AI. arXiv preprint arXiv:2012.08630 (2020)
19. Darken, R.P., Cevik, H.: Map usage in virtual environments: Orientation issues. In: Proceedings of the IEEE Conference on Virtual Reality, pp. 133–140. IEEE (1999)
20. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly* pp. 319–340 (1989)
21. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **144**(1), 114 (2015)

22. Du, Y., Tiomkin, S., Kiciman, E., Polani, D., Abbeel, P., Dragan, A.: AvE: Assistance via empowerment. *Advances in Neural Information Processing Systems* **33** (2020)
23. Eisinga, R., Te Grotenhuis, M., Pelzer, B.: The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health* **58**(4), 637–642 (2013)
24. Farrell, J.: Talk is cheap. *American Economic Review* **85**(2), 186–190 (1995)
25. Fast, E., Horvitz, E.: Long-term trends in the public perception of artificial intelligence. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31 (2017)
26. Fisher, R.A.: *Statistical Methods for Research Workers*. Oliver & Boyd (1928)
27. Fiske, S.T.: Social cognition and social perception. *Annual Review of Psychology* **44**(1), 155–194 (1993)
28. Fiske, S.T.: Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science* **27**(2), 67–73 (2018)
29. Fiske, S.T., Cuddy, A.J.C., Glick, P.: Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences* **11**(2), 77–83 (2007)
30. Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J.: A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* **82**(6), 878 (2002)
31. Foerster, J., Chen, R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., Mordatch, I.: Learning with opponent-learning awareness. In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130 (2018)
32. Gabriel, I., Ghazavi, V.: The challenge of value alignment: From fairer algorithms to AI safety. In: *The Oxford Handbook of Digital Ethics*. Oxford University Press (2021)
33. Gemp, I., McKee, K.R., Everett, R., Duéñez-Guzmán, E., Bachrach, Y., Balduzzi, D., Tacchetti, A.: D3C: Reducing the price of anarchy in multi-agent learning. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 498–506 (2022)
34. Gilad, Z., Amir, O., Levontin, L.: The effects of warmth and competence perceptions on users' choice of an AI system. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13 (2021)
35. Goodhart, C.A.E.: Problems of monetary management: The UK experience. Springer (1984)
36. Griesinger, D.W., Livingston Jr., J.W.: Toward a model of interpersonal motivation in experimental games. *Behavioral Science* **18**(3), 173–188 (1973)
37. Hamrick, J., Mohamed, S.: Levels of analysis for machine learning. In: *Bridging AI and Cognitive Science workshop at ICLR 2020* (2020)
38. Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J.L., Santoro, A.: Environmental drivers of systematicity and generalization in a situated agent. In: *International Conference on Learning Representations* (2019)
39. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation* **9**(8), 1735–1780 (1997)
40. Hutto, C., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8 (2014)
41. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. *Nature Machine Intelligence* **1**(9), 389–399 (2019)
42. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2), 263–292 (1979)
43. Karpus, J., Krüger, A., Verba, J.T., Bahrami, B., Deroy, O.: Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience* p. 102679 (2021)
44. Kelley, P.G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D., Woodruff, A.: Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, p. 627–637 (2019). DOI 10.1145/3461702.3462605
45. Kervyn, N., Fiske, S.T., Malone, C.: Brands as intentional agents framework: How perceived intentions and ability can map brand perception. *Journal of Consumer Psychology* **22**(2), 166–176 (2012)
46. Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J.T., Bernstein, M.S.: Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* **4**(CSCW2), 1–26 (2020)

47. Klatzky, R.L.: Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In: *Spatial Cognition*, pp. 1–17. Springer (1998)
48. Kollock, P.: Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology* **24**(1), 183–214 (1998)
49. Kox, E.S., Kerstholt, J.H., Huetting, T.F., de Vries, P.W.: Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems* **35**(2), 30 (2021)
50. Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent reinforcement learning in sequential social dilemmas. In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473 (2017)
51. Lerer, A., Peysakhovich, A.: Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* (2017)
52. Liebrand, W.B.G., McClintock, C.G.: The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. *European Journal of Personality* **2**(3), 217–230 (1988)
53. Likert, R.: A technique for the measurement of attitudes. *Archives of Psychology* (1932)
54. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: *Machine learning proceedings 1994*, pp. 157–163. Elsevier (1994)
55. Lockhart, E., Burch, N., Bard, N., Borgeaud, S., Eccles, T., Smaira, L., Smith, R.: Human-agent cooperation in bridge bidding. In: *Workshop on Cooperative AI at NeurIPS 2020* (2020)
56. Lord, F.M., Novick, M.R.: Statistical theories of mental test scores. Addison-Welsley Publishing (1968)
57. Marr, D.: The philosophy and the approach. In: *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press (1982)
58. McKee, K.R., Bai, X., Fiske, S.T.: Humans perceive warmth and competence in artificial intelligence. *iScience* **26**, 107256 (2023). DOI 10.1016/j.isci.2023.107256
59. McKee, K.R., Gemp, I., McWilliams, B., Duñez-Guzmán, E.A., Hughes, E., Leibo, J.Z.: Social diversity and social preferences in mixed-motive reinforcement learning. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 869–877 (2020)
60. McKee, K.R., Leibo, J.Z., Beattie, C., Everett, R.: Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems* **36**(1), 21 (2022)
61. Miller, D.: Constraints on freedom. *Ethics* **94**(1), 66–86 (1983)
62. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: *International Conference on Machine Learning*, pp. 1928–1937 (2016)
63. Murphy, R.O., Ackermann, K.A.: Social Value Orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review* **18**(1), 13–41 (2014)
64. Nagelkerke, N.J.D.: A note on a general definition of the coefficient of determination. *Biometrika* **78**(3), 691–692 (1991)
65. Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining  $r^2$  from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**(2), 133–142 (2013)
66. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
67. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. *Human Factors* **39**(2), 230–253 (1997)
68. Paulhus, D.L., Vazire, S.: The self-report method. In: *Handbook of Research Methods in Personality Psychology*, vol. 1, pp. 224–239. Guilford (2007)
69. Pe'er, E., Brandimarte, L., Samat, S., Acquisti, A.: Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* **70**, 153–163 (2017)
70. Pe'er, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E.: Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* pp. 1–20 (2021)
71. Peysakhovich, A., Lerer, A.: Consequentialist conditional cooperation in social dilemmas with imperfect information. In: *Proceedings of the 8th International Conference on Learning Representations* (2018)

72. de Quidt, J., Haushofer, J., Roth, C.: Measuring and bounding experimenter demand. *American Economic Review* **108**(11), 3266–3302 (2018)
73. Ramchurn, S.D., Wu, F., Jiang, W., Fischer, J.E., Reece, S., Roberts, S., Rodden, T., Greenhalgh, C., Jennings, N.R.: Human–agent collaboration for disaster response. *Autonomous Agents and Multi-Agent Systems* **30**, 82–111 (2016)
74. Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672–677 (2021)
75. Reeder, G.: Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological inquiry* **20**(1), 1–18 (2009)
76. Reeves, B., Hancock, J.: Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior* **1**(1) (2020). DOI 10.1037/tmb0000018
77. Russell, A.M.T., Fiske, S.T.: It's all relative: Competition and status drive interpersonal perception. *European Journal of Social Psychology* **38**(7), 1193–1201 (2008)
78. Samuelson, P.A.: A note on the pure theory of consumer's behaviour. *Economica* **5**(17), 61–71 (1938)
79. Scheunemann, M.M., Cuijpers, R.H., Salge, C.: Warmth and competence to predict human preference of robot behavior in physical human-robot interaction. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1340–1347. IEEE (2020)
80. Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., Uribe, J., Fedus, L., Metz, L., Pokorny, M., et al.: ChatGPT: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/> (2022)
81. Selten, R.: Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* **4**(1), 25–55 (1975)
82. Sevillano, V., Fiske, S.T.: Warmth and competence in animals. *Journal of Applied Social Psychology* **46**(5), 276–293 (2016)
83. Shannon, C.E.: Programming a computer for playing chess. *Philosophical Magazine* **41**(314), 256–275 (1950)
84. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**(2), 420 (1979)
85. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. *nature* **529**(7587), 484–489 (2016)
86. Singh, S.P., Barto, A.G., Chentanez, N.: Intrinsically motivated reinforcement learning. In: *Advances in Neural Information Processing Systems* (2005)
87. Siu, H.C., Pena, J.D., Chang, K.C., Chen, E., Zhou, Y., Lopez, V.J., Palko, K., Allen, R.E.: Evaluation of human-AI teams for learned and rule-based agents in Hanabi. arXiv preprint arXiv:2107.07630 (2021)
88. Slonim, R., Garbarino, E.: Increases in trust and altruism from partner selection: Experimental evidence. *Experimental Economics* **11**(2), 134–153 (2008)
89. Stanton, B., Jensen, T.: Trust and artificial intelligence. [https://tsapps.nist.gov/publication/get\\_pdf.cfm?pub\\_id=931087](https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087) (2021)
90. Strathern, M.: 'improving ratings': Audit in the British University system. *European Review* **5**(3), 305–321 (1997)
91. Strouse, D., McKee, K.R., Botvinick, M., Hughes, E., Everett, R.: Collaborating with humans without human data. *Advances in Neural Information Processing Systems* **34**, 14502–14515 (2021)
92. Tylkin, P., Radanovic, G., Parkes, D.C.: Learning robust helpful behaviors in two-player cooperative Atari environments. In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1686–1688 (2021)
93. Utz, S., Ouwerkerk, J.W., Van Lange, P.A.M.: What is smart in a social dilemma? Differential effects of priming competence on cooperation. *European Journal of Social Psychology* **34**(3), 317–332 (2004)
94. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., et al.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**(7782), 350–354 (2019)
95. Ye, C., Khalifa, A., Bontrager, P., Togelius, J.: Rotation, translation, and cropping for zero-shot generalization. In: *2020 IEEE Conference on Games*, pp. 57–64. IEEE (2020)

96. Yzerbyt, V.: The dimensional compensation model: Reality and strategic constraints on warmth and competence in intergroup perceptions. In: *Agency and Communion in Social Psychology*, pp. 126–141. Routledge (2018)
97. Zacks, J.M.: Using movement and intentions to understand simple events. *Cognitive Science* **28**(6), 979–1008 (2004)

## A Task details

This appendix provides detailed information on the implementation and parameterization of the Coins task.

We implement Coins using the DeepMind Lab2D platform [7]. Table A1 details the task settings used for agent training, as well as for the tutorial and co-play episodes in the human-agent interaction studies. Figure A2 depicts the  $11 \times 11$  room used for all co-play episodes in the interaction studies.

Following common convention in reinforcement learning research, this paper uses “step” to refer to the minimal, discrete unit of time within the environment. During each step, each player may take an action and receive rewards, and the environment can transition from one state to another.

Parameter	Agent training	Tutorial episodes	Co-play episodes
Number of players ( $n$ )	2	1	2*
Room width ( $w$ )	$\sim \mathcal{U}\{10, 15\}$	5	11
Room depth ( $d$ )	$\sim \mathcal{U}\{10, 15\}$	7	11
Coin spawn probability ( $P$ )	0.0005	0.0015	0.0005
Episode length ( $T$ )	500	1500**	300

Table A1: Parameters for the Coins task in agent training and human-agent cooperation studies. \*In Study 3, co-play episodes involve either  $n = 1$  or  $n = 2$  players, depending on whether the participant chose to play alone or with the co-player from the prior episode, respectively. \*\*Tutorial episodes terminated after participants collected five coins or after  $T = 1500$  steps (5 minutes).

### A.1 Entities and sprites

#### A.1.1 Players and coins

The two primary entities in the Coins task are players and coins (Figure A1a). We use a colorblind-friendly palette for the player and coin sprites in our experiments.

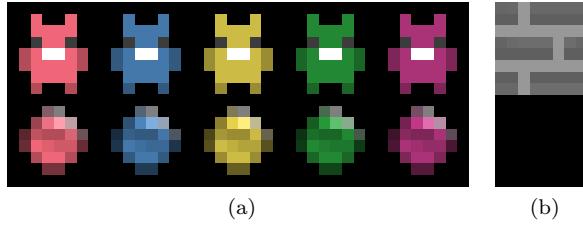


Fig. A1: Entities and sprites used in the Coins task.

Each episode of Coins involves two distinct colors sampled from the five available colors (e.g., red and blue). Each player included in an episode matches one of these two colors. On

each step, coins spawn on empty cells and cells occupied by players with probability  $P$ . When a coin spawns, it appears as one of the two colors, each with 50% probability.

Players can enter empty cells and cells with coins in them. Each player blocks other players from entering the cell they currently occupy.

### A.1.2 Miscellaneous

Walls (Figure A1b, top row) block players from moving into a cell, and thus define the boundaries and dimensions of the room. Otherwise, players are able to freely move through empty cells (Figure A1b, bottom row) in the room.

## A.2 Observations

DeepMind Lab2D supports two different visual frames of reference [47] for players:

1. *Egocentric*, meaning that players always see their avatar in an invariant position (e.g., centered) in their visual input
2. *Allocentric*, meaning that the world itself remains in an invariant position (e.g., centered) in players' visual input

Egocentric frames of reference promote generalization for reinforcement learning agents [38, 95]. Consequently, we provide egocentric observations for our agents (Figure A3a). The agent observation spans five cells in each direction (forward, backward, left, and right) from the agent. Given the sprite dimensions ( $8 \times 8 \times 3$ , with the last dimension reflecting RGB channels), agents observed their environment through a  $88 \times 88 \times 3$  window.

In pilot tests of our study interface, human players reported being confused by the changes in their visual observation when playing with an egocentric reference frame. In combination with prior evidence that allocentric reference frames can support human navigation in virtual environments [19], this feedback prompted us to provide human players with an allocentric reference frame for the environment (Figure A3b). As human players move their avatar around the Coins gridworld, the room stays centered within their visual input while their avatar moves around the room.

## A.3 Actions

Players can take one of the following five actions each step:

1. **No-op**: Makes no change to the player's position.
2. **Move up**: Moves the player up one cell.
3. **Move down**: Moves the player down one cell.
4. **Move left**: Moves the player left one cell.
5. **Move right**: Moves the player right one cell.

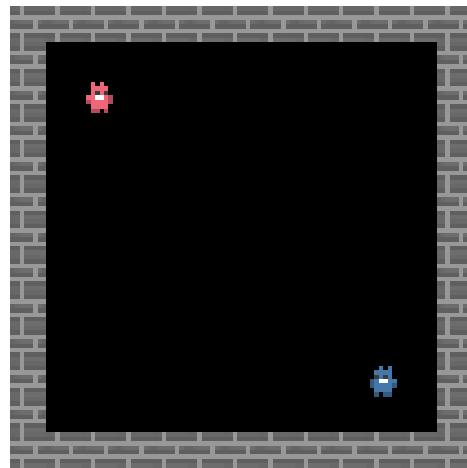


Fig. A2: A **11 × 11** room, as used in all co-play episodes in human-agent cooperation studies.

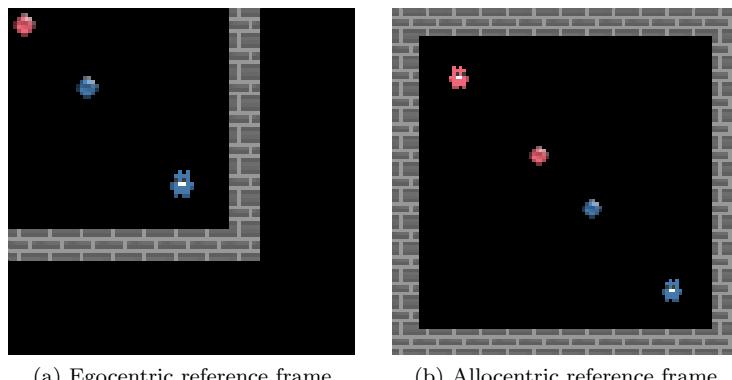


Fig. A3: Observations for the Coins task, depicting different frames of reference for the same example game state. (a) Agent players observe Coins with an egocentric reference frame. (b) Human players observe Coins with an allocentric reference frame.

## B Agent details

This appendix provides detailed information on the architecture, parameterization, and training of the artificial agents that we study.

### B.1 Agent architecture

We build an advantage actor-critic (A2C) agent with two added algorithmic components (Figure A4). The Social Value Orientation (SVO) component—integrated for training, evaluation, and co-play—recomputes the environment reward signal received by the agent before its use by the critic. As detailed in Section B.2, the  $\theta$  parameter guides this computation. The trembling-hand component—incorporated for evaluation and co-play—sits between the actor and the environment, replacing each action emitted by the policy with a random action with probability  $\epsilon$ .

### B.2 Social Value Orientation

McKee et al. [59] introduce and define the SVO algorithmic component across three descriptive levels [37, 57]:

1. At the *computational level*, McKee et al. propose SVO as a mechanism “redefin[ing] self-interest to incorporate the interests of a broader group”.
2. At the *algorithmic level*, McKee et al. introduce “reward angles” as a method of representing distributions of reward over the self and the group.
3. At the *implementation level*, McKee et al. offer a penalty-based approach that assigns pseudoreward to SVO agents based on the divergence between their target reward angle and the realized reward angle, in combination with a weight parameter  $w$ . Eq. (3) in [59] defines this implementation.

Here we introduce a new implementation of SVO that retains the computational and algorithmic levels as described above, but replaces the divergence-penalization method with an approach based on vector projection (Figure A4). Within the Markov game framework described in [59], we define the vector-projection approach with a function  $U_i$  to be maximized by agent  $i$ :

$$U_i(s, o_i, a_i) = r_i \cdot \cos \theta + \bar{r}_{-i} \cdot \sin \theta \quad (1)$$

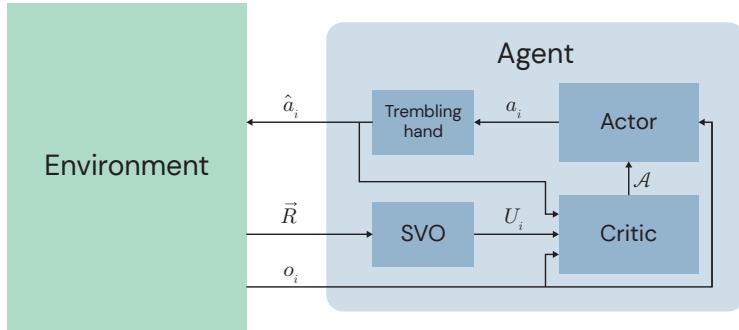
In the Markov game framework,  $s$  is the current environmental state;  $o_i$  is the observation received by agent  $i$ ;  $a_i$  is the action selected by agent  $i$  through a policy  $\pi(a_i|o_i)$ ;  $r_i$  is the environmental reward received by agent  $i$ , contained within the state reward vector  $\vec{R} = (r_1, \dots, r_n)$  for all  $n$  agents in the environment;  $\theta$  is the Social Value Orientation for player  $i$  (representing player  $i$ 's target distribution of reward among group members); and  $\bar{r}_{-i}$  is a statistic summarizing the rewards of all other group members from  $\vec{R}$ . For these experiments, we choose to define  $\bar{r}_{-i}$  as the arithmetic mean:  $\bar{r}_{-i} = \frac{1}{n-1} \sum_{j \neq i} r_j$ , where  $n$  is the group size. This formulation of SVO parallels the model of human preferences described by Griesinger and Livingston [36].

This utility function leaves most of its inputs unused, aside from information contained in the group reward vector  $\vec{R}$ . Consequently, the vector-projection approach may also be referenced as  $U_i(\vec{R})$ , as in Figure A4.

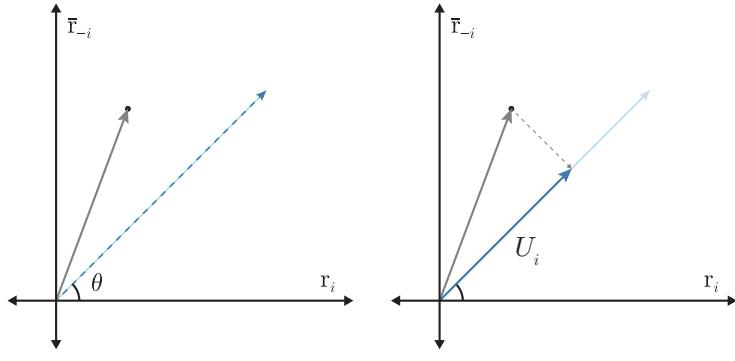
### B.3 Training protocol

Agents trained in a distributed set of environments running in parallel. Each agent trained using one GPU, for a total of 5e7 environment steps across 150 environments. Our distributed framework checkpointed agents every 15 minutes during training.

We compute training curves by evaluating each agent checkpoint in 100 self-play episodes, then calculating the mean and standard deviation of total coin collections (Figure A5), mismatching coin collections (Figure A6), and collective return (Figure A7) for each set of evaluation episodes.



(a) Agent architecture.



(b) Vector projection and SVO.

Fig. A4: Agent details. (a) In the Markov game framework, agents select actions to take on the basis of observations and rewards they receive from the environment. We construct an advantage actor-critic (A2C) agent with two algorithmic modifications: a Social Value Orientation (SVO) component that takes the group reward vector  $\vec{R}$  as input and computes utility  $U_i$  for the critic's use; and a “trembling hand” component that changes the actor's output  $a_i$  with some probability  $\epsilon$ , producing a modified action  $\hat{a}_i$ . The agent otherwise operates as usual, with the actor initially selecting an action given the agent's observation  $o_i$ , and the critic computing an advantage estimate  $\mathcal{A}$  from the agent's observation  $o_i$ , its utility  $U_i$ , and the agent's modified action  $\hat{a}_i$ . For more information on A2C, see [62]. (b) The SVO component computes  $U_i$  through vector projection. On the left, the black dot represents the reward vector given by the environment and defined by  $r_i$  and  $\bar{r}_{-i}$ . The dotted blue line represents a vector along the agent's SVO angle. On the right,  $U_i$  is the magnitude of the blue vector computed by projecting the environmental reward vector on the dotted blue line. See Eq. 1 for precise formulation.

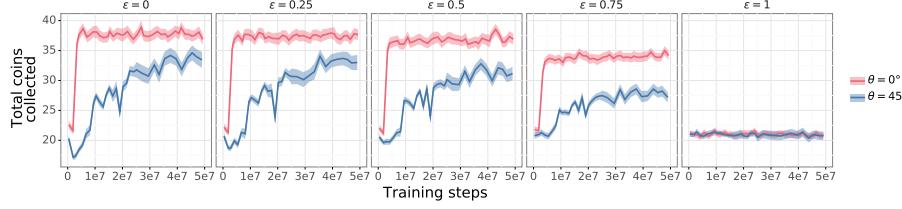


Fig. A5: Effect of  $\epsilon$  on total coin collection over training. Error bands represent 95% confidence intervals over 100 evaluation episodes at each training checkpoint.

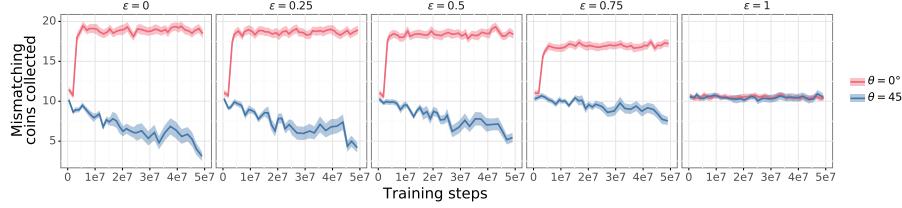


Fig. A6: Effect of  $\epsilon$  on mismatching-coin collection over training. Error bands represent 95% confidence intervals over 100 evaluation episodes at each training checkpoint.

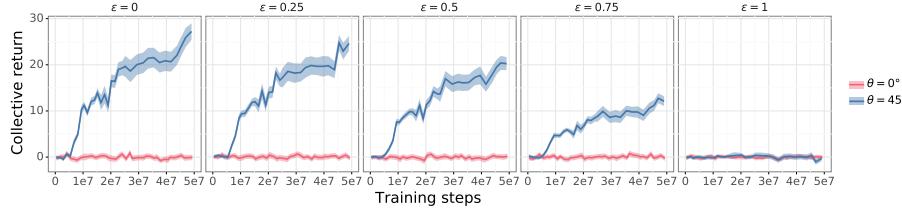


Fig. A7: Effect of  $\epsilon$  on collective return over training. Error bands represent 95% confidence intervals over 100 evaluation episodes at each training checkpoint.

## C Human-agent interaction studies

This appendix provides detailed information on the design of the human-agent interaction studies and on the statistical analyses conducted in each study.

### C.1 Study details

All three studies received a favorable opinion from the Human Behavioural Research Ethics Committee at DeepMind (#19/004) and were approved by the Institutional Review Board for Human Subjects at Princeton University (#11885).

Each study applied the same inclusion criteria during recruitment on Prolific: residence in the United States, prior completion of at least 20 studies on Prolific, and an approval rating of 95% or more.

We include screenshots showing how Study 1 unfolded for each participant:

1. Read general instructions on the study and gameplay (Figure A8).
2. Play practice episode (Figure A9).
3. Read instructions on game rules and co-players (Figures A10 and A11).
4. Play co-play episode with Co-player A and answer questions about perceptions (Figure A12).
5. Play co-play episode with Co-player B and answer questions about perceptions and preferences (Figure A13).
6. Repeat steps 3 and 4 for another 10 episodes.
7. Transition to post-task questionnaire (Figure A14).

Figure A15 shows an example open-ended impression question from the post-task questionnaire.

During gameplay, a “ticker” at the top of the screen displayed the three most recent coin collection events. The study interface did not provide participants with cumulative metrics (e.g., score or coin collections).

In all three studies, the practice episode instructed participants to collect five coins. The practice episode ended after the participant collected five coins or five minutes elapsed, whichever occurred first.

In Studies 1 and 2, we generated the sequence of co-players for each participant by listing the 12 possible pairwise combinations of agents, randomly shuffling the order of the pairs, and then randomly shuffling the order of the agents within each pair.

Study 2 closely resembled Study 1, aside from identifying the bonus earned per point as \$0.02 (Figures A10a and A11h) and using the reward values from the shifted incentive structure (Table 2) when explaining the rules for matching and mismatching coin collections (Figures A11b-A11e).

Study 3 largely preserved the instructions preceding the first co-play episode (Figures A8-A11), though the wording on certain screens changed slightly to refer to a single upcoming episode rather than multiple episodes (e.g., “the next round” rather than “each round”). The study branched more substantially starting just before the first co-play episode, as shown in Figures A16 and A17. After playing a single co-play episode, participants read a description of their partner choice for the following episode. They played the partner-choice episode as they chose (that is, either alone or with the co-player from the previous episode).

### C.2 Analytic details

For our statistical analyses, we rely on several regression methods to help examine our data. In situations where our data consist of independent observations, we employ standard linear and logistic regressions to analyze the relationships between variables. For analyses involving repeated measurements collected from each participant, we leverage generalized linear mixed-effect models (GLMMs). These models offer flexibility to model different types of outcome measures (e.g., preferences reported on a bounded scale) and account for non-independent observations.

In addition to fitting regression models, we leverage the ANOVA method (standing for “analysis of variance”). ANOVAs [26] allow us to test whether changing the value of an independent variable (e.g., an agent hyperparameter) significantly affects the value of a specified dependent variable (e.g., human perceptions of warmth or competence). Each ANOVA is summarized with an  $F$ -statistic, two subsetted values on the  $F$ -statistic (the between-groups degrees of freedom and the within-groups degrees of freedom), and a  $p$ -value.

### C.2.1 Study 1

We expect the social perception questions to exhibit suitable psychometric properties: namely, participants should offer relatively consistent judgments across repeated interactions with a given agent; participants should vary more in their judgments between traits than within a particular trait; and the two items within each composite measure (“warm” and “well-intentioned” for warmth, “competent” and “intelligent” for competence) should correlate consistently.

As reported in the main text (Table 4), participants made highly consistent judgments for a given agent on a given trait. A mixed ANOVA indicated that social perceptions varied more between traits than within a particular trait,  $F_{4,3744} = 96.2, p < 0.001$ . Finally, both the composite warmth measure ( $\rho = 0.93$ ) and the composite competence measure ( $\rho = 0.92$ ) exhibit high internal consistency, as measured by the Spearman-Brown formula.

A two-way mixed ANOVA modeled the effects of  $\theta$ ,  $\epsilon$ , and their interaction on warmth evaluations. Figure 5a depicts the main effect of  $\theta$  on perceived warmth (that is, marginalized over  $\epsilon$  values), and Figure A18a visualizes the full two-way interaction. The following means and standard deviations help describe main effects by marginalizing over all other variables. Participants perceived  $\theta = 45^\circ$  agents ( $m = 3.52, sd = 1.40$ ) as significantly warmer than  $\theta = 0^\circ$  agents ( $m = 1.75, sd = 1.31$ ),  $F_{1,1108} = 1006.8, p < 0.001$ . The  $\epsilon$  parameter also exerted a significant effect on warmth judgments, with  $\epsilon = 0.5$  agents ( $m = 2.73, sd = 1.33$ ) perceived as significantly warmer than  $\epsilon = 0$  agents ( $m = 2.55, sd = 1.49$ ),  $F_{1,1108} = 9.2, p = 0.002$ . Finally, the effect of the interaction between the  $\theta$  and  $\epsilon$  parameters was significant,  $F_{1,1108} = 56.7, p < 0.001$ .

A two-way mixed ANOVA modeled the effects of  $\theta$ ,  $\epsilon$ , and their interaction on competence judgments. Figure 5b depicts the main effect of  $\epsilon$  on competence evaluations (that is, marginalized over  $\theta$  values), and Figure A18b visualizes the full two-way interaction. The following means and standard deviations help describe main effects by marginalizing over all other variables. Participants judged  $\epsilon = 0$  agents ( $m = 3.53, sd = 1.28$ ) as significantly more competent than  $\epsilon = 0.5$  agents ( $m = 3.00, sd = 1.18$ ),  $F_{1,1108} = 70.6, p < 0.001$ . The  $\theta$  parameter also exerted a significant effect on competence evaluations,  $F_{1,1108} = 25.8, p < 0.001$ . Participants perceived  $\theta = 0^\circ$  agents ( $m = 3.42, sd = 1.29$ ) as significantly more competent than  $\theta = 45^\circ$  agents ( $m = 3.10, sd = 1.21$ ). The effect of the interaction between the  $\theta$  and  $\epsilon$  parameters was not significant,  $F_{1,1108} = 1.5, p = 0.22$ .

A sequence of fractional response models evaluated the predictive value of the hypothesized predictors for participants’ stated pairwise preferences. We construct each model through a GLMM with a binomial link function, normalizing the full preference scale with zero and one as its endpoints. To ensure model comparability, all fractional response models predict the exact same outcome variable.

The first GLMM predicted stated preferences from a joint intercept, a fixed effect for the identities of the agents being compared, and a random effect for the participant. The underlying identities of the agents exerted a significant influence on the participant’s stated pairwise preferences,  $\chi^2(11) = 572.8, p < 0.001$ . Figure A19 shows mean pairwise preferences calculated for each pair of agent identities. The effect of the joint intercept was not significant,  $p = 0.32$ . The model achieves an AIC of 1661.9 and  $R_m^2 = 0.362$ .

The second GLMM predicted stated preferences from a joint intercept, a fixed effect for the difference in scores the participant received when playing with each of the agents being compared, and a random effect for the participant. The difference in scores earned by the participant significantly affected stated pairwise preferences, OR = 1.12, 95% CI [1.11, 1.13],  $p < 0.001$ . The effect of the joint intercept was not significant,  $p = 0.12$ . The model achieves an AIC of 1697.2 and  $R_m^2 = 0.363$ .

The third GLMM predicted stated preferences from a joint intercept, a fixed effect for the difference in perceived warmth for each of the agents being compared, a fixed effect for the difference in perceived competence for each of the agents, and a random effect for the participant. The difference in the perceived warmth of the agents significantly influenced stated

pairwise preferences,  $OR = 2.23$ , 95% CI [2.08, 2.40],  $p < 0.001$ . The difference in the perceived competence of the agents significantly altered stated pairwise preferences,  $OR = 0.78$ , 95% CI [0.73, 0.84],  $p < 0.001$ . The effect of the joint intercept was not significant,  $p = 0.23$ . The model achieves an AIC of 1610.2 and  $R^2_m = 0.435$ .

A final GLMM directly compared the predictiveness of social perceptions against that of the score that participants received. For this model, we standardized the predictors, allowing direct comparison between the model coefficients. The GLMM predicted pairwise preferences from: a joint intercept; standardized variables for perceived warmth, perceived competence, and score; and a random intercept for participant. All three fixed effects were significant within this model (Figure A20). One standard deviation of change in the warmth variable (difference in perceived warmth between two agents) had the largest standardized effect on pairwise preferences,  $OR = 3.02$ , 95% CI [2.54, 3.60],  $p < 0.001$ . One standard deviation in score had the next largest effect,  $OR = 2.13$ , 95% CI [1.82, 2.50],  $p < 0.001$ . One standard deviation in the competence variable (difference in perceived competence between two agents) had the effect with the smallest magnitude,  $OR = 0.85$ , 95% CI [0.75, 0.96],  $p = 0.012$ . The effect of the joint intercept was significant,  $OR = 1.18$ , 95% CI [1.03, 1.36],  $p = 0.012$ .

Finally, a linear model predicted post-game impression sentiment from an intercept, perceived warmth, and perceived competence. Perceived warmth significantly correlated with impression sentiment,  $\beta = 0.13$ , 95% CI [0.09, 0.16],  $p < 0.001$ . Perceived competence did not exhibit a significant relationship with sentiment,  $p = 0.24$ . The effect of the intercept was not significant,  $p = 0.52$ .

### C.2.2 Study 2

As an initial test, we sought to validate the effect of shifting the incentive structure on participants' rewards. A linear mixed-effects model predicted participant reward from a categorical variable representing the study number, with Study 1 as the reference level. As expected, the study variable exerted a significant effect on participant reward,  $\beta = 27.3$ , 95% CI [26.5, 28.0],  $p < 0.001$ .

As in Study 1, we expect that the social perception questions will exhibit suitable psychometric properties. Table 6 shows the high consistency with which participants evaluated a given agent on a given trait. A mixed ANOVA showed that social perceptions varied more between traits than within a particular trait,  $F_{4,4650} = 88.5$ ,  $p < 0.001$ . Finally, both the composite warmth measure ( $\rho = 0.92$ ) and the composite competence measure ( $\rho = 0.91$ ) exhibit high internal consistency, as measured by the Spearman-Brown formula.

A two-way mixed ANOVA modeled the effects of  $\theta$ ,  $\epsilon$ , and their interaction on warmth perceptions. Figure A21a depicts the main effect of  $\theta$  on warmth judgments (that is, marginalized over  $\epsilon$  values), and Figure A22a visualizes the full two-way interaction. The following means and standard deviations help describe main effects by marginalizing over all other variables. Participants assessed  $\theta = 45^\circ$  agents ( $m = 3.48$ ,  $sd = 1.18$ ) as significantly warmer than  $\theta = 0^\circ$  agents ( $m = 1.79$ ,  $sd = 1.00$ ),  $F_{1,1086} = 981.9$ ,  $p < 0.001$ . The  $\epsilon$  parameter also exerted a significant effect on warmth evaluations,  $F_{1,1086} = 20.3$ ,  $p = 0.002$ . Participants perceived  $\epsilon = 0.5$  agents ( $m = 2.76$ ,  $sd = 1.23$ ) as significantly warmer than  $\epsilon = 0$  agents ( $m = 2.52$ ,  $sd = 1.51$ ). Finally, the effect of the interaction between the  $\theta$  and  $\epsilon$  parameters was significant,  $F_{1,1086} = 123.4$ ,  $p < 0.001$ .

A two-way mixed ANOVA modeled the effects of  $\theta$ ,  $\epsilon$ , and their interaction on competence evaluations. Figure A21b depicts the main effect of  $\epsilon$  on competence evaluations (that is, marginalized over  $\theta$  values), and Figure A22b visualizes the full two-way interaction. The following means and standard deviations help describe main effects by marginalizing over all other variables. Participants judged  $\epsilon = 0$  agents ( $m = 3.49$ ,  $sd = 1.27$ ) as significantly more competent than  $\epsilon = 0.5$  agents ( $m = 2.95$ ,  $sd = 1.09$ ),  $F_{1,1086} = 76.0$ ,  $p < 0.001$ . The  $\theta$  parameter also exerted a significant effect on competence evaluations, with  $\theta = 0^\circ$  agents ( $m = 3.35$ ,  $sd = 1.22$ ) perceived as significantly more competent than  $\theta = 45^\circ$  agents ( $m = 3.09$ ,  $sd = 1.20$ ),  $F_{1,1086} = 16.9$ ,  $p < 0.001$ . Finally, the effect of the interaction between the  $\theta$  and  $\epsilon$  parameters was not significant,  $F_{1,1086} = 2.1$ ,  $p = 0.15$ .

A sequence of fractional response models assessed the predictive value of the hypothesized predictors for participants' stated pairwise preferences. We construct each model through a GLMM with a binomial link function, normalizing the full preference scale with zero and one as its endpoints. To ensure model comparability, all fractional response models predict the exact same outcome variable.

The first GLMM predicted stated preferences from a joint intercept, a fixed effect for the identities of the agents being compared, and a random effect for the participant. The underlying identities of the agents significantly affected the participant's stated pairwise preferences,  $\chi^2(11) = 549.7, p < 0.001$ . Figure A23 shows mean pairwise preferences calculated for each pair of agent identities. The effect of the joint intercept was not significant,  $p = 0.15$ . The model achieves an AIC of 1608.7 and  $R^2_m = 0.403$ .

The second GLMM predicted stated preferences from a joint intercept, a fixed effect for the difference in scores the participant received when playing with each of the agents being compared, and a random effect for the participant. The difference in scores earned by the participant significantly influenced stated pairwise preferences, OR = 1.06, 95% CI [1.06, 1.07],  $p < 0.001$ . The effect of the joint intercept was not significant,  $p = 0.81$ . The model achieves an AIC of 2049.4 and  $R^2_m = 0.214$ .

The third GLMM predicted stated preferences from a joint intercept, a fixed effect for the difference in perceived warmth for each of the agents being compared, a fixed effect for the difference in perceived competence for each of the agents, and a random effect for the participant. The difference in the perceived warmth of the agents significantly altered stated pairwise preferences, OR = 2.63, 95% CI [2.42, 2.89],  $p < 0.001$ . The difference in the perceived competence of the agents significantly affected stated pairwise preferences, OR = 0.82, 95% CI [0.76, 0.88],  $p < 0.001$ . The effect of the joint intercept was not significant,  $p = 0.99$ . The model achieves an AIC of 1510.4 and  $R^2_m = 0.496$ .

A final GLMM directly compared the predictiveness of social perceptions against that of the score that participants received. For this model, we standardized the predictors, allowing direct comparison between the model coefficients. The GLMM predicted pairwise preferences from: a joint intercept; standardized variables for perceived warmth, perceived competence, and score; and a random intercept for participant. All three fixed effects were significant within this model (Figure A24). One standard deviation of change in the warmth variable (difference in perceived warmth between two agents) had the largest standardized effect on pairwise preferences, OR = 5.56, 95% CI [4.74, 6.74],  $p < 0.001$ . One standard deviation in score had the next largest effect, OR = 1.50, 95% CI [1.31, 1.71],  $p < 0.001$ . One standard deviation in the competence variable (difference in perceived competence between two agents) had the effect with the smallest magnitude, OR = 0.80, 95% CI [0.70, 0.91],  $p < 0.001$ . The effect of the joint intercept was not significant,  $p = 0.49$ .

Finally, a linear model predicted post-game impression sentiment from an intercept, perceived warmth, and perceived competence. Perceived warmth significantly correlated with impression sentiment,  $\beta = 0.12$ , 95% CI [0.09, 0.15],  $p < 0.001$  (Figure 11a). Perceived competence also significantly correlated with sentiment,  $\beta = 0.04$ , 95% CI [0.00, 0.08],  $p = 0.037$  (Figure 11b). The effect of the intercept on impression sentiment was significant,  $\beta = -0.15$ , 95% CI [-0.26, -0.05],  $p = 0.005$ .

### C.2.3 Study 3

In Study 3, each participant interacted with a single agent, and thus only provided a single measurement point for warmth judgments and competence judgments. As a result, the ANOVAs and regressions reported below do not incorporate random effects for participants.

An ANOVA indicated that social perceptions varied more between traits than within a particular trait,  $F_{3,1200} = 54.9, p < 0.001$ . Both the composite warmth measure ( $\rho = 0.85$ ) and the composite competence measure ( $\rho = 0.86$ ) exhibit high internal consistency, as measured by the Spearman-Brown formula.

A two-way ANOVA modeled the effects of  $\theta$ ,  $\epsilon$ , and their interaction on warmth evaluations. Figure A25a depicts the main effect of  $\theta$  on perceived warmth (that is, marginalized over  $\epsilon$  values), and Figure A26a visualizes the full two-way interaction. The following means and standard deviations help describe main effects by marginalizing over all other variables. Participants perceived  $\theta = 45^\circ$  agents ( $m = 3.14, sd = 1.22$ ) as significantly warmer than  $\theta = 0^\circ$  agents ( $m = 1.79, sd = 1.09$ ),  $F_{1,297} = 103.4, p < 0.001$ . The  $\epsilon$  parameter did not exert a significant effect on warmth judgments,  $F_{1,297} = 0.2, p = 0.62$ . Finally, the effect of the interaction between the  $\theta$  and  $\epsilon$  parameters was significant,  $F_{1,297} = 5.3, p = 0.022$ .

A two-way ANOVA modeled the effects of  $\theta$ ,  $\epsilon$ , and their interaction on competence judgments. Figure A25b depicts the main effect of  $\epsilon$  on competence evaluations (that is, marginalized over  $\theta$  values), and Figure A26b visualizes the full two-way interaction. The following

means and standard deviations help describe main effects by marginalizing over all other variables. Participants judged  $\epsilon = 0$  agents ( $m = 3.75$ ,  $sd = 1.15$ ) as significantly more competent than  $\epsilon = 0.5$  agents ( $m = 2.95$ ,  $sd = 1.23$ ),  $F_{1,297} = 35.3$ ,  $p < 0.001$ . The  $\theta$  parameter did not exert a significant effect on competence evaluations,  $F_{1,297} = 3.8$ ,  $p = 0.05$ . Finally, the effect of the interaction between the  $\theta$  and  $\epsilon$  parameters was not significant,  $F_{1,297} = 1.2$ ,  $p = 0.26$ .

A sequence of logistic regressions evaluated the predictive value of the hypothesized predictors for participants' revealed preferences.

The first logistic regression predicted revealed preferences from an intercept and the identity of the agent. The underlying identity of the agent significantly affected the participant's revealed preferences,  $\chi^2(3) = 45.4$ ,  $p < 0.001$ . The effect of the intercept on revealed preferences was significant, OR = 1.88, 95% CI [1.18, 3.07],  $p = 0.009$ . The model achieves an AIC of 372.5 and  $R^2 = 0.188$ .

The second logistic regression predicted revealed preferences from an intercept and the score earned by the participant when playing with the agent. The score earned by the participant exerted a significant influence on revealed preferences, OR = 1.06, 95% CI [1.03, 1.08],  $p < 0.001$ . The effect of the intercept was also significant, OR = 0.16, 95% CI [0.08, 0.31],  $p < 0.001$ . The model achieves an AIC of 390.5 and  $R^2 = 0.101$ .

The third logistic regression predicted revealed preferences from an intercept, perceived warmth, and perceived competence. Perceived warmth significantly predicted revealed preferences, OR = 2.10, 95% CI [1.69, 2.65],  $p < 0.001$ . Perceived competence did not significantly predict revealed preferences,  $p = 0.88$ . The effect of the intercept was not significant,  $p = 0.23$ . The model achieves an AIC of 356.2 and  $R^2 = 0.242$ .

A final logistic regression directly compared the predictiveness of social perceptions against that of the score that participants received. For this model, we standardized the predictors, allowing direct comparison between the model coefficients. The model predicted revealed preferences from an intercept and standardized variables for perceived warmth, perceived competence, and score (Figure A27). The effects of perceived warmth and of score were significant within the model. One standard deviation of change in perceived warmth had the larger effect on revealed preferences, OR = 2.27, 95% CI [1.65, 3.16],  $p < 0.001$ . One standard deviation in score had a smaller effect, OR = 1.47, 95% CI [1.09, 1.98],  $p = 0.011$ . The effect of perceived competence was not significant,  $p = 0.44$ . The effect of the intercept was significant, OR = 1.18, 95% CI [1.03, 1.36],  $p = 0.006$ .

Finally, a linear model predicted post-game impression sentiment from an intercept, perceived warmth, and perceived competence. Perceived warmth significantly correlated with impression sentiment,  $\beta = 0.14$ , 95% CI [0.10, 0.18],  $p < 0.001$  (Figure 14a). Perceived competence also significantly correlated with sentiment,  $\beta = 0.04$ , 95% CI [0.00, 0.08],  $p = 0.041$  (Figure 14b). The effect of the intercept on impression sentiment was significant,  $\beta = -0.15$ , 95% CI [-0.25, -0.05],  $p = 0.002$ .

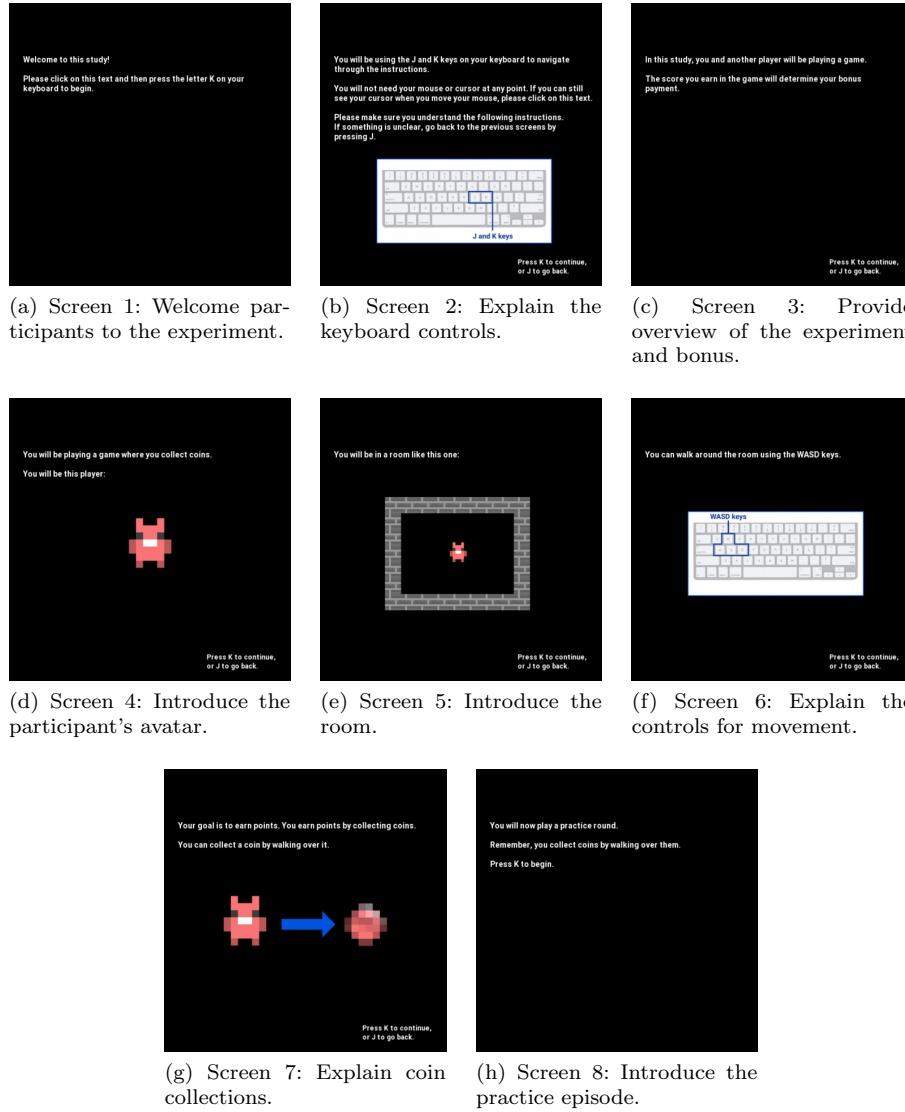


Fig. A8: Screenshots of instruction screens in Study 1.

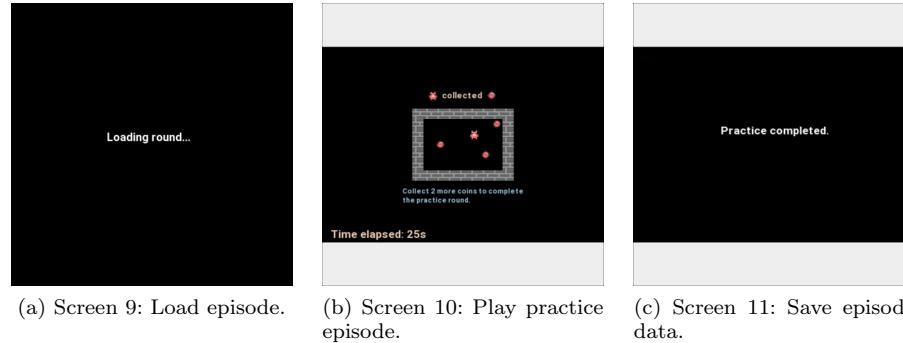


Fig. A9: Screenshots of practice episode in Study 1.

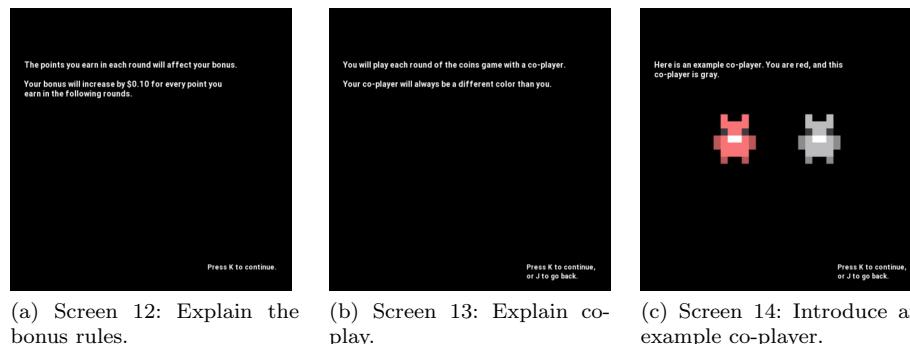


Fig. A10: Screenshots of instruction screens in Study 1.

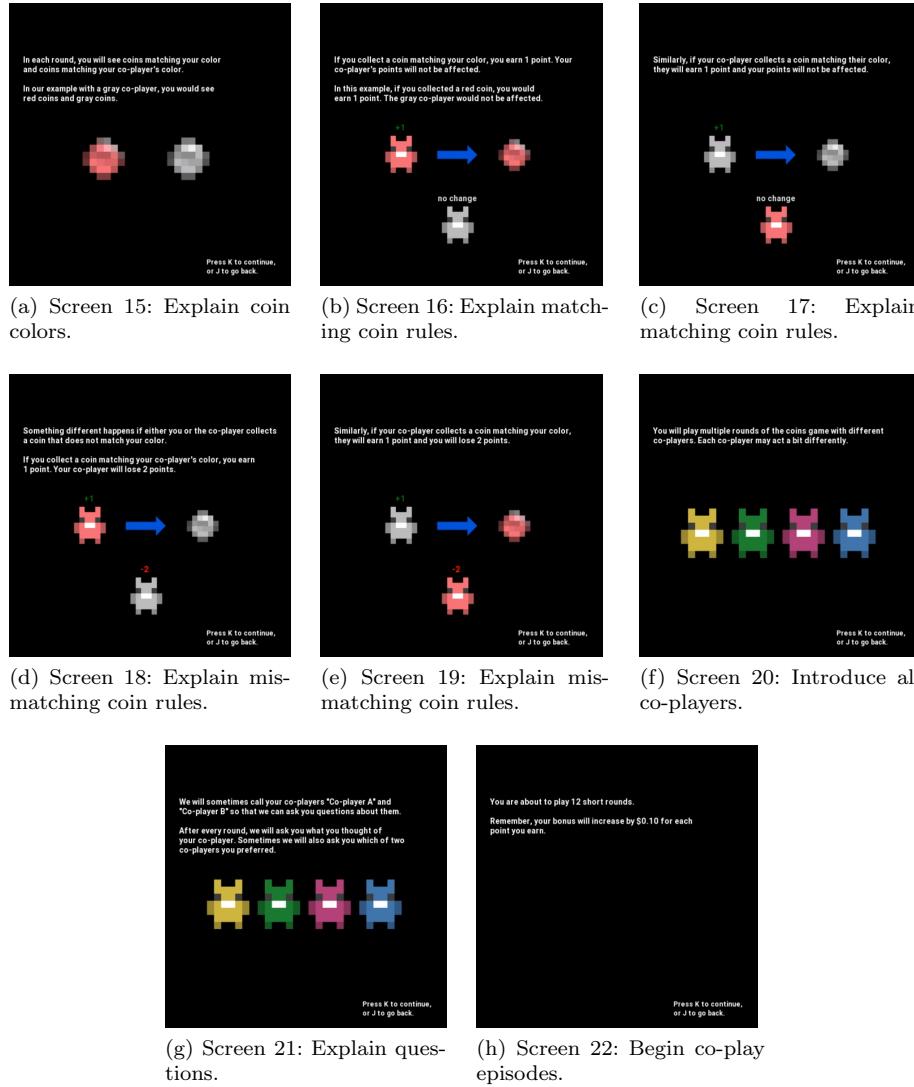


Fig. A11: Screenshots of instruction screens in Study 1.

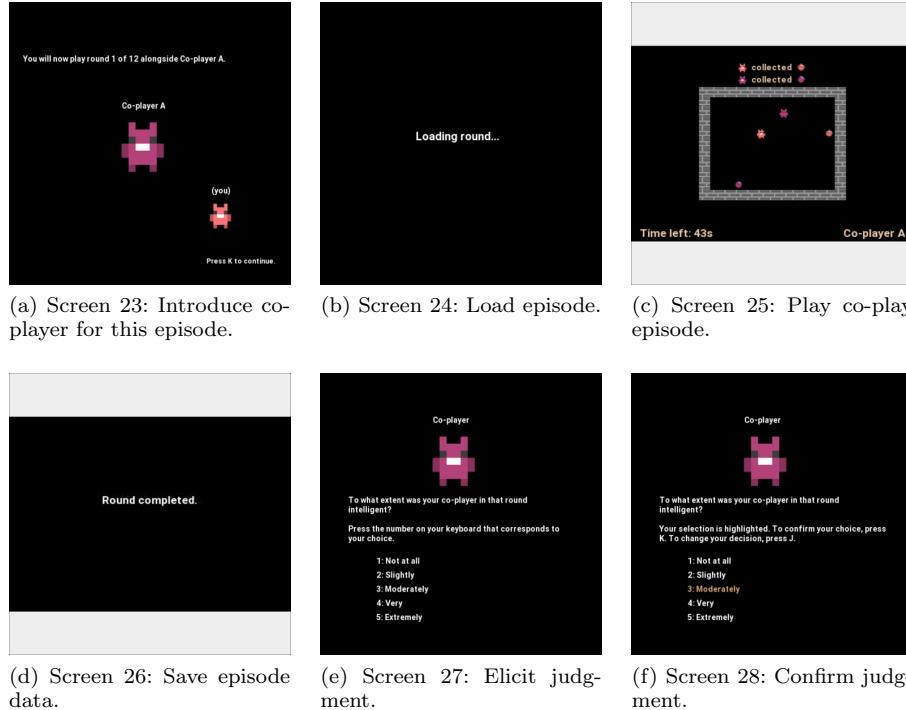


Fig. A12: Screenshots of an “A” co-play episode and subsequent questions on perceptions in Study 1. (e) and (f) show the screens for eliciting and confirming one judgment; the study repeats these screens for all four judgments, randomizing the sequence of judgments for each episode.

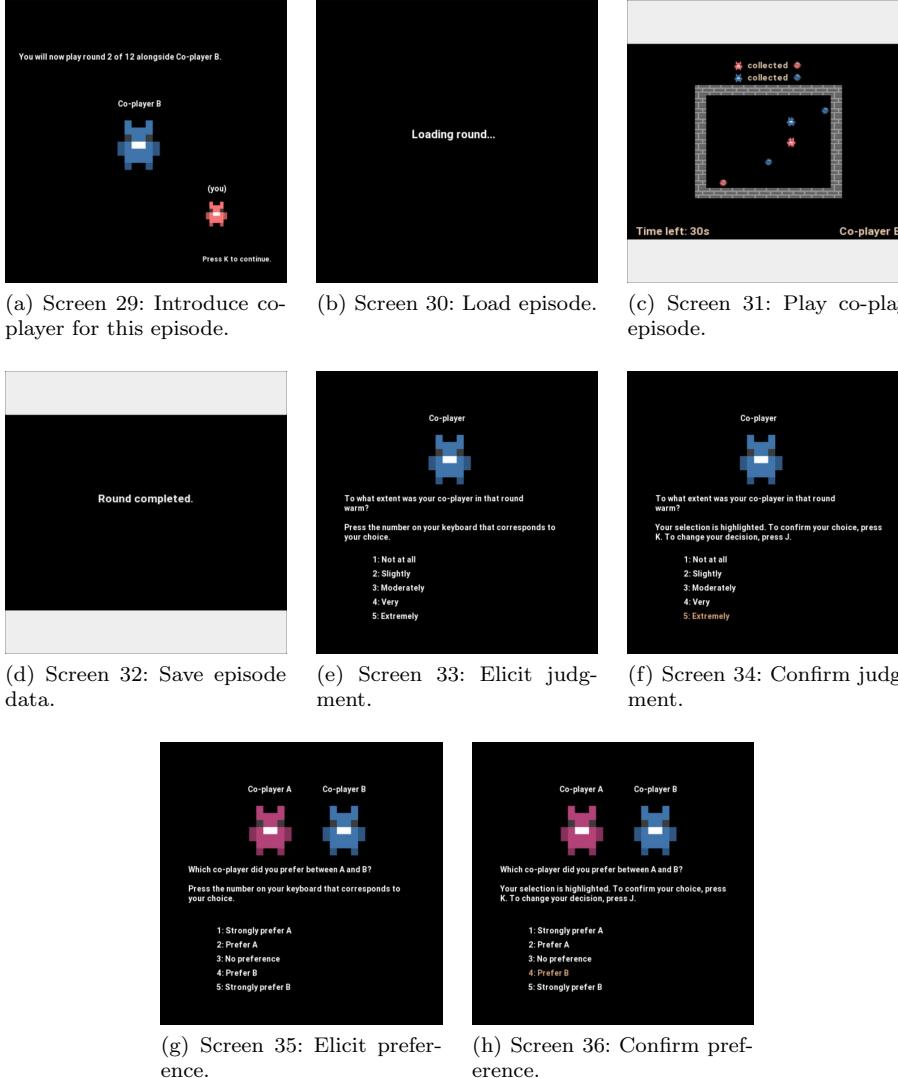
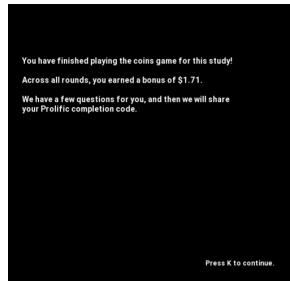


Fig. A13: Screenshots of a “B” co-play episode and subsequent questions on perceptions and preferences in Study 1. (e) and (f) show the screens for eliciting and confirming one judgment; the study repeats these screens for all four judgments, randomizing the sequence of judgments for each episode.



(a) Screen 37: Confirm bonus and transition to post-task questionnaire.

Fig. A14: Screenshot of pre-questionnaire page in Study 1.

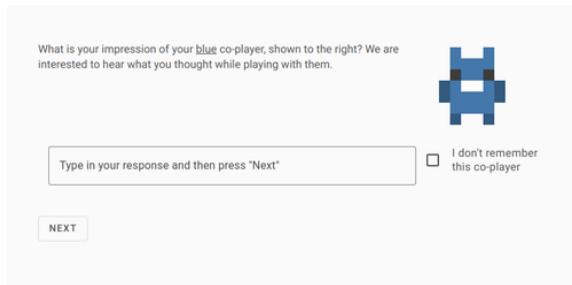


Fig. A15: Screenshot of open-ended impression question from post-task questionnaire in Study 1.

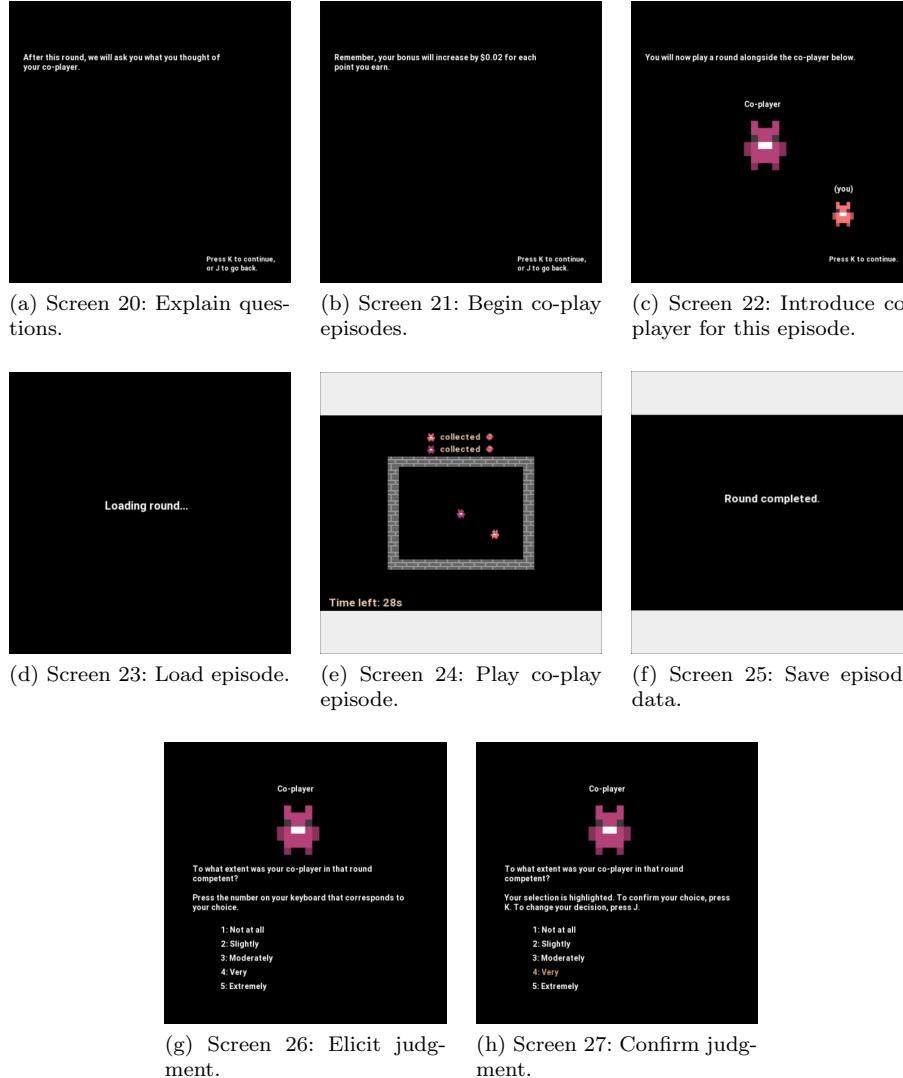


Fig. A16: Screenshots of instructions, co-play episode, and subsequent questions on perceptions in Study 3. (g) and (h) show the screens for eliciting and confirming one judgment; the study repeats these screens for all four judgments, randomizing the sequence of judgments for each episode.

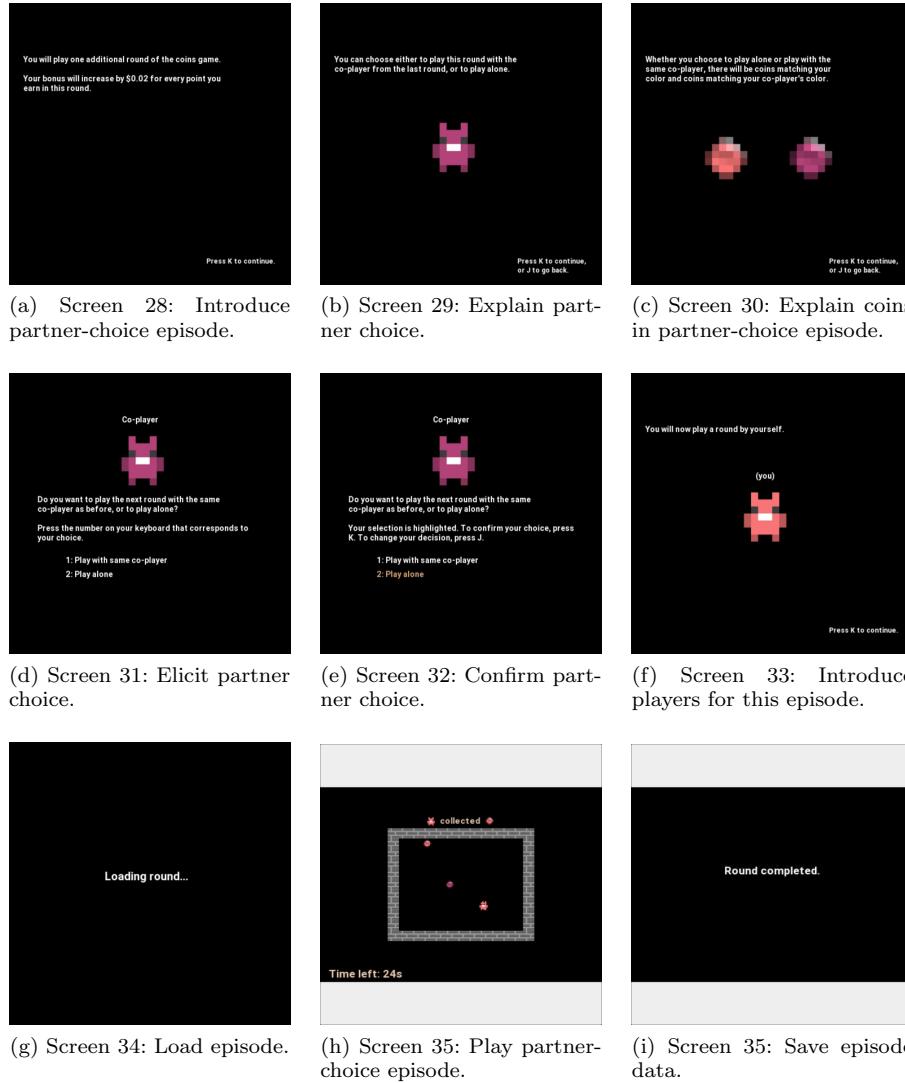


Fig. A17: Screenshots of instructions, partner choice question, and partner-choice episode in Study 3.

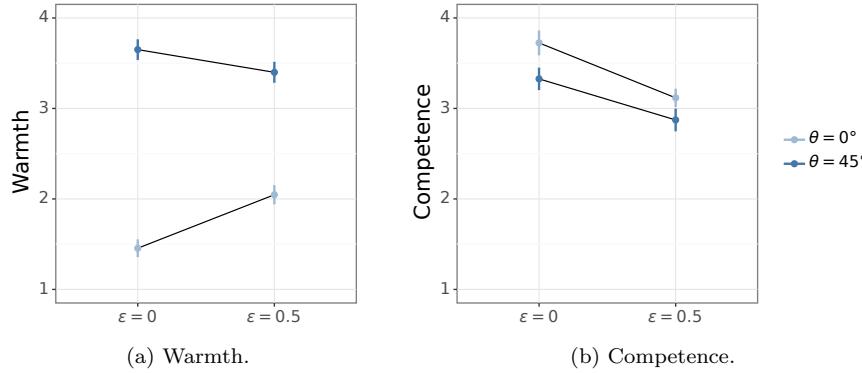


Fig. A18: Interaction between the SVO and trembling-hand components on social perceptions in Study 1. Error bars reflect 95% confidence intervals.

		$\theta = 45^\circ, \epsilon = 0$	$\theta = 45^\circ, \epsilon = 0.5$	$\theta = 0^\circ, \epsilon = 0.5$	$\theta = 0^\circ, \epsilon = 0$
		-0.02	1.17	1.50	
		0.02	1.07	1.45	
$\theta = 0^\circ$	$\epsilon = 0.5$	-1.17	-1.07		0.99
	$\epsilon = 0$	-1.50	-1.45	-0.99	

Fig. A19: Pairwise preferences for different agents in Study 1. Cell values represent self-reported preferences for the row agent over the column agent.

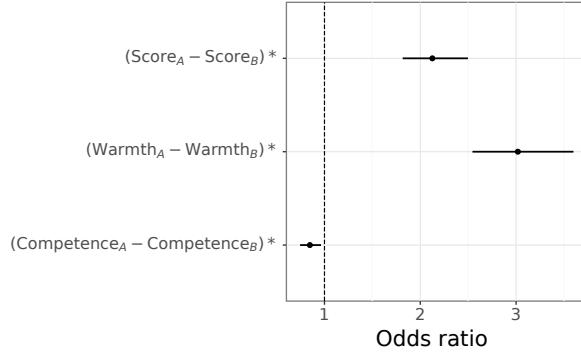


Fig. A20: Odds ratios in Study 1. Asterisks indicate that predictors are centered and normalized by one standard deviation to permit fair comparison between their effect sizes.

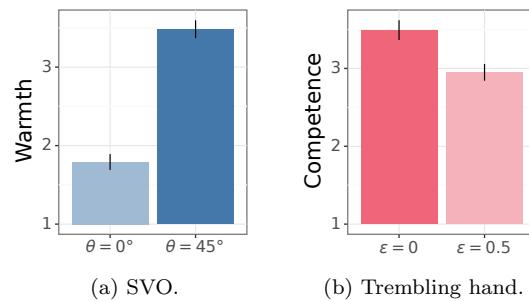


Fig. A21: Main effects of the SVO and trembling-hand components on social perceptions in Study 2. (a) An agent's SVO parameter significantly affected warmth perceptions,  $p < 0.001$ . (b) The trembling-hand component significantly influenced competence evaluations,  $p < 0.001$ . Error bars reflect 95% confidence intervals.

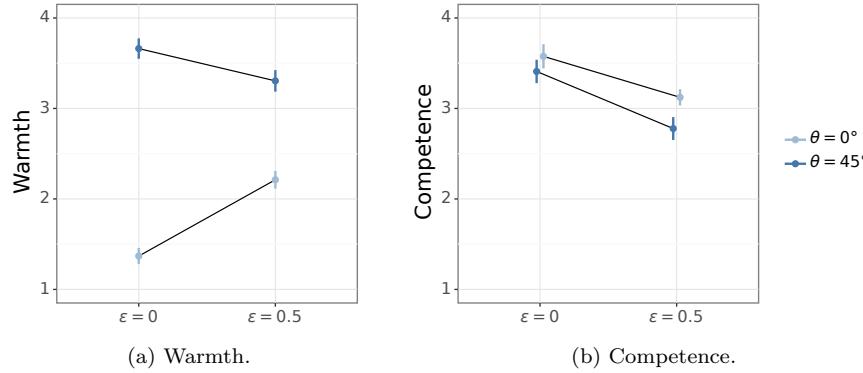


Fig. A22: Interaction between the SVO and trembling-hand components on social perceptions in Study 2. Error bars reflect 95% confidence intervals.

	$\theta = 45^\circ, \varepsilon = 0$	$\theta = 45^\circ, \varepsilon = 0.5$	$\theta = 0^\circ, \varepsilon = 0.5$	$\theta = 0^\circ, \varepsilon = 0$
$\theta = 45^\circ, \varepsilon = 0$		0.26	1.12	1.62
$\theta = 45^\circ, \varepsilon = 0.5$	-0.26		0.96	1.46
$\theta = 0^\circ, \varepsilon = 0.5$	-1.12	-0.96		0.87
$\theta = 0^\circ, \varepsilon = 0$	-1.62	-1.46	-0.87	

Fig. A23: Pairwise preferences for different agents in Study 2. Cell values represent self-reported preferences for the row agent over the column agent.

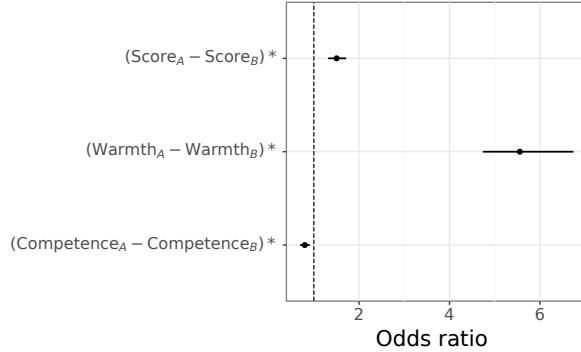


Fig. A24: Odds ratios in Study 2. Asterisks indicate that predictors are centered and normalized by one standard deviation to permit fair comparison between their effect sizes.

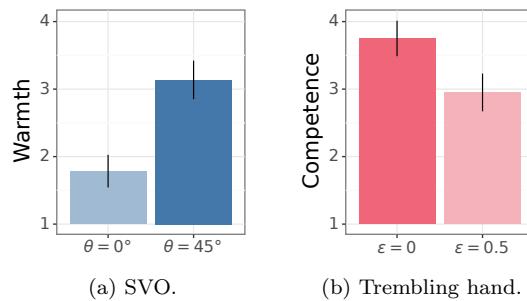


Fig. A25: Main effects of the SVO and trembling-hand components on social perceptions in Study 3. (a) The SVO component significantly altered warmth judgments,  $p < 0.001$ . (b) Similarly, an agent's trembling-hand parameterization significantly influenced perceived competence,  $p < 0.001$ . Error bars reflect 95% confidence intervals.

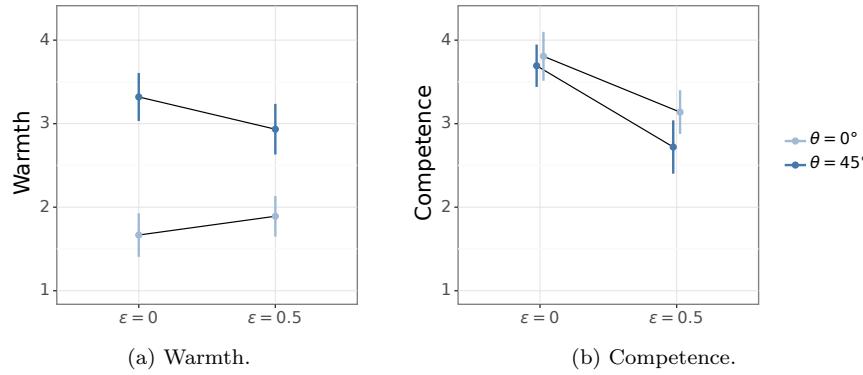


Fig. A26: Interaction between the SVO and trembling-hand components on social perceptions in Study 3. Error bars reflect 95% confidence intervals.

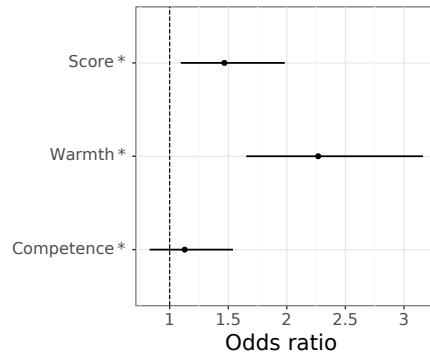


Fig. A27: Odds ratios in Study 3. Asterisks indicate that predictors are centered and normalized by one standard deviation to permit fair comparison between their effect sizes.