Warmth and Competence in Human-Agent Cooperation

Kevin R. McKee
DeepMind
London, United Kingdom
kevinrmckee@deepmind.com

Xuechunzi Bai
Princeton University
Princeton, New Jersey, United States
xb2@princeton.edu

Susan T. Fiske
Princeton University
Princeton, New Jersey, United States
sfiske@princeton.edu

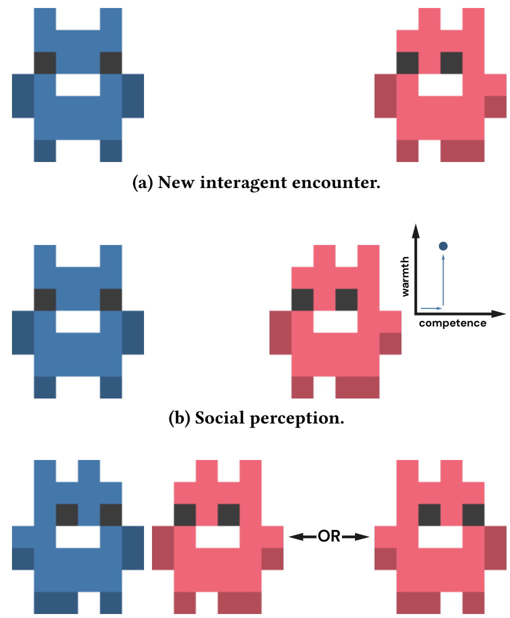
ABSTRACT

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 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 round 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

ACM Reference Format:
Kevin R. McKee, Xuechunzi Bai, and Susan T. Fiske. 2022. Warmth and Competence in Human-Agent Cooperation. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 10 pages.



(a) New interagent encounter.

(b) Social perception.

(c) Partner choice.

Figure 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).

1 INTRODUCTION

Trust is central to the development and deployment of artificial intelligence (AI) [34, 74]. However, many members of the public harbor doubts and concerns about the trustworthiness of AI [13, 18, 22, 37]. This presents a pressing issue for cooperation between humans and AI agents [15].

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 [12, 46, 76, 77]. 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 [72, 76]. 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, 24, 25, 47]. These perceptions help individuals "make sense of [other actors] in order to guide their own actions and interactions" [23] (Figure 1). The competence dimension aligns with the established focus on performance and score in machine learning research [8]: How

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

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, 26]. 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.

Until now, experiments evaluating human views on agents have relied on stated preferences, often by directly asking participants which of two agents they preferred as a partner [19, 72, 76]. Such self-report methods can be insightful tools for research [57]. However, they are vulnerable to experimenter demand [17] and exhibit limited ecological validity. In this paper, we overcome these challenges by eliciting *revealed preferences* [65]: 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 [4, 5, 10, 73]. In incentivized experiments, partner-choice measures mitigate experimenter demand [17]. 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 [6, 16, 56]. Finally, 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 [7, 34, 50] into human-agent interaction research.

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

(1) Demonstrates how reinforcement learning can be used to train human-compatible agents for a temporally and spatially extended mixed-motive game.
(2) Measures both stated and revealed preferences, introducing a partner choice framework for the latter.
(3) Examines how fundamental social perceptions affect stated and revealed preferences over agents, above and beyond traditional objective metrics.

Readers can find the appendix (including additional figures and tables) in the extended version of this paper on arXiv [48].

2 METHODS

2.1 Task

Coins [27, 29, 42, 60] (Figure 2) is a mixed-motive Markov game [45] played by $n = 2$ players. 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 coin matches one of the two players in color. On each step of the game, players can stand still or move around the room.

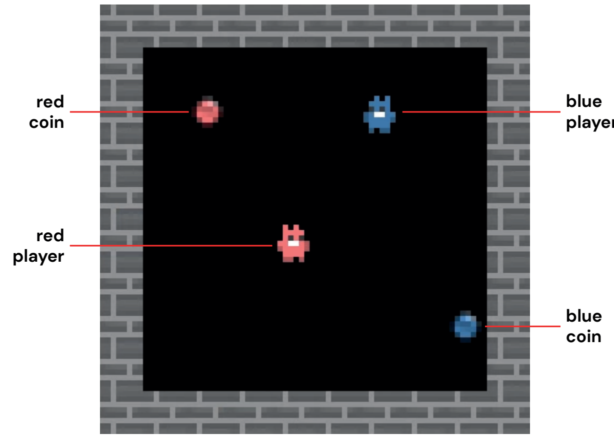


Figure 2: Screenshot of gameplay in Coins. Two players move around a small room and collect colored coins. Coins randomly appear in the room over time. Players receive reward from coin collections depending on the match or mismatch between their color and the coin color.

The goal of the game is to earn reward by collecting coins. Players pick up coins by stepping onto them. Coin collections generate reward 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, 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 [40]: players can pursue selfish goals or prosocial goals. Second, relative to matrix games like the Prisoner's Dilemma, Coins is temporally and spatially extended [41]: players can efficiently or inefficiently achieve their goals. We hypothesize that these two features offer sufficient affordance for an observer to infer other players' intentions and their effectiveness at enacting their intentions [9, 62, 79].

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. Overall, our study design is agnostic to the algorithmic implementation of the agents being evaluated. For this paper, the agents learn using the advantage actor-critic (A2C) algorithm [51]. The neural network consists of a convolutional module, a fully connected module, an LSTM with contrastive predictive coding [32, 55], and linear readouts for policy and value. Agents train for 5×10^7 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 [49], an algorithmic module inspired by psychological models of human prosocial preferences [31, 43, 52], into our agents. The SVO component parameterizes each agent with θ, representing a target distribution over their reward and the reward of other agents in their environment. SVO agents are intrinsically motivated [71] to optimize for task rewards that align with their parameterized target θ. 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" [14, 67] component to the agents for evaluation and co-play. The trembling-hand module replaces each action selected by the agent with probability ϵ. 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 $\epsilon = 0.5$.

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

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

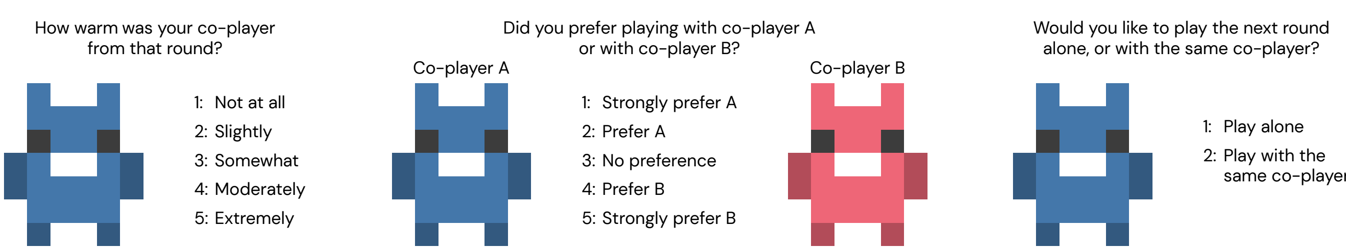
Table 3: Predictions for social perception (warmth and competence) as a function of agent hyperparameters (Social Value Orientation θ and trembling hand ϵ).

2.3 Study design for human-agent studies

We recruited participants from Prolific [58, 59] 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 studies. Appendix C contains further 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 approaches these constructs using the canonical payoff structure for Coins [42] and an established self-report framework for eliciting (stated) preferences [76]. We next 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? Given that humans respond more strongly to losses than to commensurate gains [35], we tested participants' partner choices under a shifted incentive structure with all non-negative outcomes (Table 2). To ensure that any differences in results stem from the switch from stated to revealed preferences, we break this question into two studies, changing a single variable at a time between studies. Study 2 uses the same stated-preference approach as Study 1, but incorporates the offset incentive structure. Study 3 then elicits revealed preferences in place of stated preferences.

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%



(a) Social perception. **(b) Stated preferences.** **(c) Revealed preferences.**

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

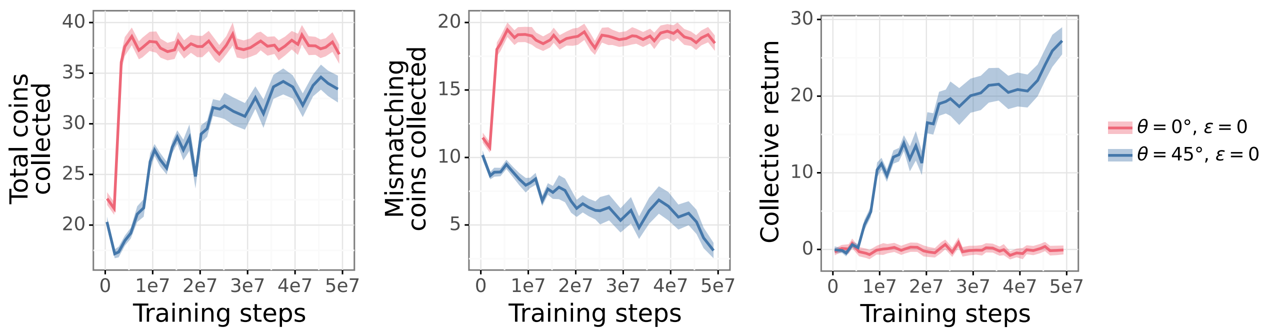


Figure 4: Performance metrics over agent training. Selfish agents quickly learned to collect coins, but did not learn to avoid mismatches. As a result, collective return hovered around zero. Prosocial agents exhibited slower learning and collected fewer coins on average, but also learned to avoid mismatching 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.

female, 51.5% male; $m_{\text{age}} = 34$, $sd_{\text{age}} = 13$). The study employed a within-participant 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 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). 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 debrief questionnaire. The questionnaire first solicited open-ended responses about each of 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; $m_{\text{age}} = 34$, $sd_{\text{age}} = 12$). The study employed the same within-participant design as Study 1, with one primary change: participants and agents played Coins under the shifted incentive structure.

The study instructed participants that they would receive \$0.02 for every point earned during the remaining 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. 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; $m_{\text{age}} = 33$, $sd_{\text{age}} = 11$). In contrast with the preceding studies, Study 3 employed a between-participant 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, 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 debrief 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.

3 RESULTS

3.1 Agent training

Figure 4 displays coin collections and score over the course of agent training. The training curves for $\theta = 0°$ agents closely resemble those from previous studies [42]: 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 ($\theta = 45°$) 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 ϵ values degrade performance. Total coin collections decrease with increasing ϵ for both selfish and prosocial agents. Higher levels of ϵ cause prosocial agents to become less discerning at avoiding mismatching coins, and consequently produce lower levels of collective return.

3.2 Human-agent studies

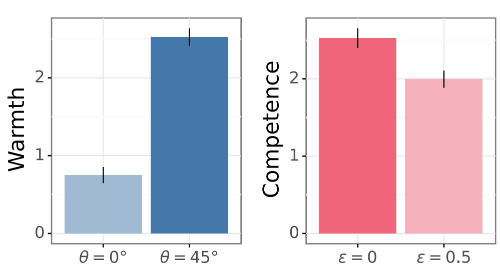
3.2.1 Study 1. Participants played with each agent three times during the study, evaluating the relevant agent after each round of play. Participants did not make judgments at random; their responses were highly consistent across their interactions with each agent (Table 4). At the same time, participants were not submitting vacuous appraisals. 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). ICC ranges from 0 to 1, with higher values indicating greater consistency.

Following standard practice in social perception research [26, 44], we combine individual judgments into composite warmth and competence measures for subsequent analysis. Both composite measures exhibit high scale reliability as measured by the Spearman-Brown formula [20], 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).



(a) SVO. **(b) Trembling hand.**

Figure 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.

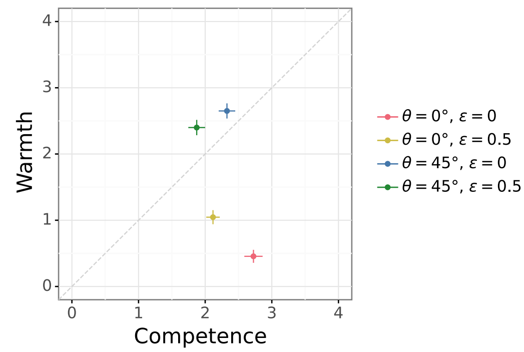


Figure 6: Overall pattern of perceived warmth and competence in Study 1. Error bars reflect 95% confidence intervals.

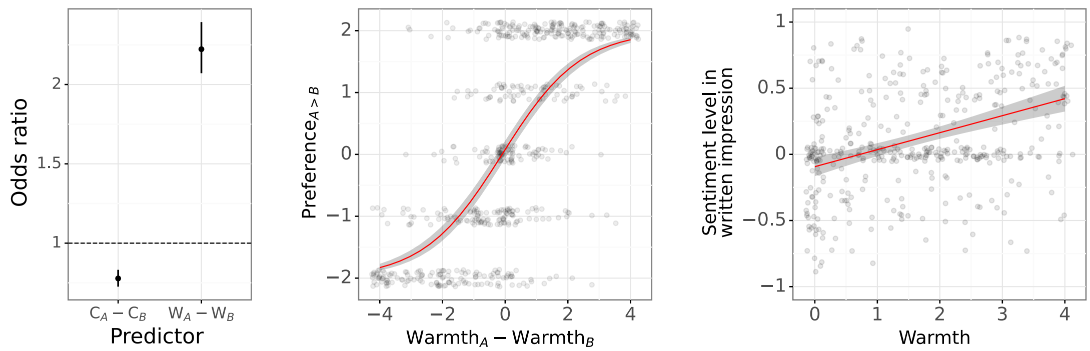
Stated preferences. How well do participants' perceptions predict subjective preferences, relative to predictions made based on objective score? We fit competing fractional response models regressing self-reported preferences on score and social perception, respectively, and then compared model fit using the Akaike information criterion (AIC) [2] and Nakagawa's R^2 [54]. We fit an additional baseline model using algorithm identities (i.e., which two agents participants were choosing between) as a predictor.

The model leveraging algorithm identities and the model leveraging participant scores both accounted for a large amount of variance in subjective preferences. Participants exhibited a clear pattern of preferences across the four agents (Figure A19). In pairwise comparison, participants favored the $\theta = 45°$ agents over both $\theta = 0°$ agents, and preferred the $\theta = 0°$, $\epsilon = 0.5$ agent over the $\theta = 0°$, $\epsilon = 0$ 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 odds ratio OR = 1.12, 95% CI [1.11, 1.13], $p < 0.001$.

Predictor	AIC	R^2_m
Algorithm identities	1661.9	0.362
Participant score	1697.2	0.363
Social perception	**1611.2**	**0.436**

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.

Nonetheless, knowing participants' judgments generates substantially better predictions of their preferences than the alternatives (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, 2.41], $p < 0.001$ (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, 0.83],



(a) Odds ratios from model predicting stated preferences.

(b) Effect of perceived warmth on stated preferences.

(c) Effect of perceived warmth on impression sentiment.

Figure 7: Relationship between social perception and subjective preferences in Study 1. Participants' evaluations of warmth and competence improved predictions of stated preferences for different co-players, above and beyond participant score. (a) and (b) depict odds ratios and preference predictions from a fractional-response regression, respectively. (c) plots sentiment predictions from a linear regression. Error bars and bands represent 95% confidence intervals.

$p < 0.001$. As a further demonstration of the predictive power of participants' social perceptions, the effect of warmth and competence remains significant when entered in a regression alongside score (Figure A20). Social perception thus improves model fit above and beyond that provided by score alone.

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) [33]. 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 $[0.04, 0.22]$, $p = 0.004$ (Figure 7). In contrast, competence did not exhibit a significant relationship with sentiment, $p = 0.45$.

Anecdotally, participants expressed a wide range of emotions while describing their co-players. The $\theta = 45°$ agents often evoked contrition and guilt. For example, one participant stated, "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."

Participants discussed the $\theta = 0°$ agents, on the other hand, with anger and frustration. One participant reported, "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. Overall, the 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 2). 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 8). 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. Participants reported a clear pattern of preferences across the agents (Figure A23). In pairwise comparison, participants favored the $\theta = 45°$ agents over the $\theta = 0°$ agents, and the $\theta = 0°$, $\epsilon = 0.5$ agent over the $\theta = 0°$, $\epsilon = 0$ agent.

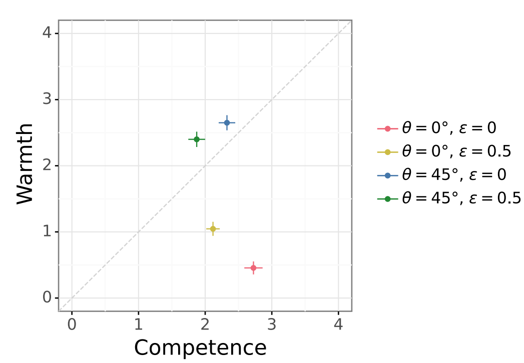


Figure 8: Overall pattern of perceived warmth and competence in Study 2. Error bars depict 95% confidence intervals.

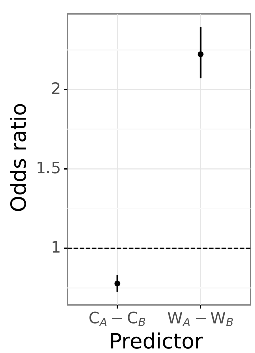
The model with participant score as the sole predictor performed considerably worse than it did in Study 1 (Table 6, middle row). Still, it captured the same pattern as before: the higher a participant scored with co-player A relative to co-player B, the greater their preferences for co-player A, OR = 1.06, 95% CI [1.06, 1.07], $p < 0.001$.

Predictor	AIC	R^2_m
Co-player identities	1608.7	0.403
Participant score	2049.4	0.214
Social perception	**1509.8**	**0.499**

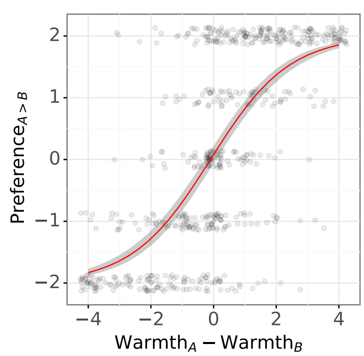
Table 6: Metrics for fractional response models predicting preferences in Study 2. Lower values of AIC and higher values of R^2_m indicate stronger fits.

Participants' perceptions again serve as a better foundation for preference predictions than game score or the identity of the specific algorithms they encountered (Table 6, bottom row, and Figure 9a). 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, 2.87], $p < 0.001$ (Figure 9b). 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 [0.75, 0.87], $p < 0.001$. Perceived warmth and competence remain significant predictors of participants' preferences when entered in a regression alongside score, improving model fit above and beyond that provided by score alone (Figure A24).

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



(a) Odds ratios from model predicting preferences.



(b) Effect of perceived warmth on preferences.

Figure 9: Relationship between social perception and subjective preferences in Study 2, as modeled through fractional-response regression. Participants' evaluations of warmth and competence improved predictions of stated preferences for different co-players, above and beyond participant score. Error bars and bands reflect 95% confidence intervals.

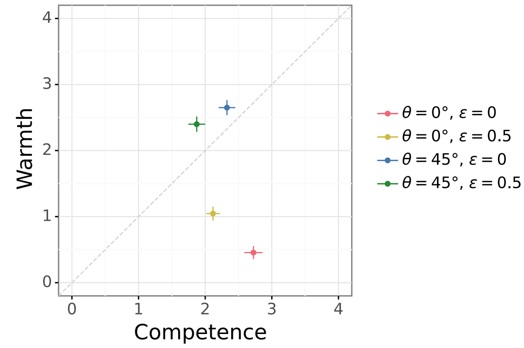


Figure 10: Overall pattern of perceived warmth and competence in Study 3. Error bars reflect 95% confidence intervals.

positively they tended to describe that co-player, $\beta = 0.11$, 95% CI [0.03, 0.19], $p = 0.008$ (Figure A25a). Breaking from the prior study, perceptions of competence exhibited a similar effect on impression sentiment: the more competent an agent seemed, the more positively participants described them, $\beta = 0.04$, 95% CI [−0.02, 0.09], $p = 0.037$ (Figure A25b).

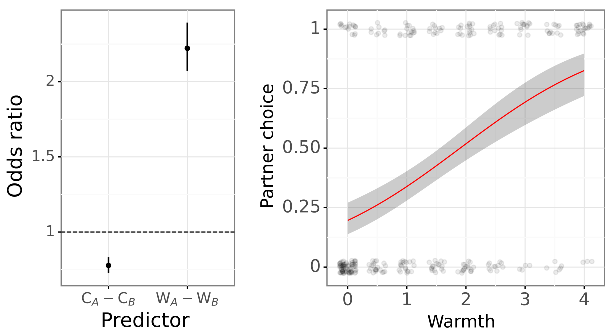
3.2.3 Study 3. Our final study tested whether the relationship between social perceptions and subjective preferences translates to a revealed-preference setting. Does social perception continue to predict preferences when individuals face a partner choice?

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 asseen in Studies 1 and 2 (Figure 10). Participants perceived high-SVO agents as significantly warmer than low-SVO agents, $F_{1,297} = 103.4$, $p < 0.001$ (Figure A26a). Similarly, steady agents came across as significantly more competent than trembling agents, $F_{1,297} = 49.6$, $p < 0.001$ (Figure A26b).

Predictor	AIC	R^2
Co-player identities	372.5	0.188
Scores	390.5	0.101
Social perception	**357.9**	**0.243**

Table 7: Metrics for logistic models predicting partner choice in Study 3. Lower values of AIC and higher values of R^2 indicate stronger fits.

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 [53]. Social perception again proves a better predictor than score or even co-player identity (Table 7). The warmer a co-player appeared to participants, the more likely participants were to play another round with them, OR = 2.1, 95% CI [1.69, 2.65], $p < 0.001$ (Figure 11b). There was no significant relationship between perceived competence and partner choice, $p = 0.88$. Perceived warmth and competence remain significant predictors of participants' preferences when entered in a



(a) Odds ratios from model predicting partner choice.

(b) Effect of perceived warmth on partner choice.

Figure 11: Relationship between social perception and subjective preferences in Study 3, as modeled through logistic regression. Participants' evaluations of warmth and competence improved predictions of revealed preferences for different co-players, above and beyond participant score. Error bars and bands indicate 95% confidence intervals.

regression alongside score, improving model fit above and beyond that provided by score alone (Figure A28).

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 A29a). Despite the lack of correspondence between perceived competence and partner choice, perceptions of competence exhibited a similar effect on 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 A29b).

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. Agents elicited varying perceptions of warmth and competence when interacting with humans. Objective features like game score predict interactants' preferences over different agents. However, preference predictions substantially improve by taking into account people's perceptions. 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. Follow-up research will help explore the mechanisms underlying (and the robustness of) this effect.

These results reinforce the generality of warmth and competence. Perceptions of warmth and competence structure impressions of other humans [64], as well as impressions of non-human actors including animals [68], corporations [38], and robots [63, 66]. In combination with recent studies of human-agent interactions in consumer decision-making contexts [30, 39] and the Prisoner's Dilemma [47], 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 [11, 69, 70, 78]. We follow recent calls to move AI research beyond competition and toward cooperation [15]. Most interaction research on deep reinforcement learning focuses on pure common-interest games such as Overcooked [12, 76] and Hanabi [72], 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 [3]. At the same time, it conflicts with recent work arguing that humans are "keen to exploit benevolent AI" [36]. Future 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" [21]. 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., [61]). In a similar vein, subjective preferences may serve as a valuable objective for optimization.[1] Future studies can investigate the viability of such an approach.

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" [28]. In his extensive discussion of value alignment, Gabriel [30] 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.

ACKNOWLEDGMENTS

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.

[1] Of course, this approach carries risk. As recognized by Charles Goodhart and Marilyn Strathern, "when a measure becomes a target, it ceases to be a good measure" [75].

REFERENCES

[1] Andrea E. Abele, Naomi Ellemers, Susan T. Fiske, Alex Koch, and Vincent Yzerbyt. 2021. Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review* 128, 2 (2021), 290.

[2] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723.

[3] Ruben T. Azevedo, Maria Serena Panasiti, Rosita Maglio, and Salvatore Maria Aglioti. 2018. Perceived warmth and competence of others shape voluntary deceptive behaviour in a morally relevant setting. *British Journal of Psychology* 109, 1 (2018), 25–44.

[4] Pat Barclay and Robb Willer. 2007. Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences* 274, 1610 (2007), 749–753.

[5] Nicolas Baumard, Jean-Baptiste André, and Dan Sperber. 2013. A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences* 36, 1 (2013), 59–78.

[6] Anne Beaudry and Alain Pinsonneault. 2010. The other side of acceptance: Studying the direct and indirect effects of emotions on information technology use. *Management Information Systems Quarterly* (2010), 689–710.

[7] Isaiah Berlin. 1969. Four essays on liberty. (1969).

[8] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).

[9] Sarah-Jayne Blakemore and Jean Decety. 2001. From the perception of action to the understanding of intention. *Nature Reviews Neuroscience* 2, 8 (2001), 561–567.

[10] Martin Brown, Armin Falk, and Ernst Fehr. 2004. Relational contracts and the nature of market interactions. *Econometrica* 72, 3 (2004), 747–780.

[11] Murray Campbell, A. Joseph Hoane Jr., and Feng-hsiung Hsu. 2002. Deep Blue. *Artificial Intelligence* 134, 1-2 (2002), 57–83.

[12] Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-AI coordination. *Advances in Neural Information Processing Systems* 32 (2019), 5174–5185.

[13] Stephen Cave and Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1, 2 (2019), 74–78.

[14] Fiery Cushman, Anna Dreber, Ying Wang, and Jay Costa. 2009. Accidental outcomes guide punishment in a "trembling hand" game. *PLOS One* 4, 8 (2009), e6699.

[15] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

[16] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly* (1989), 319–340.

[17] Jonathan de Quidt, Johannes Haushofer, and Christopher Roth. 2018. Measuring and bounding experimenter demand. *American Economic Review* 108, 11 (2018), 3266–3302.

[18] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[19] Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca Dragan. 2020. AvE: Assistance via empowerment. *Advances in Neural Information Processing Systems* 33 (2020).

[20] Rob Eisinga, Manfred Te Grotenhuis, and Ben Pelzer. 2013. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health* 58, 4 (2013), 637–642.

[21] Joseph Farrell. 1995. Talk is cheap. *American Economic Review* 85, 2 (1995), 186–190.

[22] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[23] Susan T. Fiske. 1993. Social cognition and social perception. *Annual Review of Psychology* 44, 1 (1993), 155–194.

[24] Susan T. Fiske. 2018. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science* 27, 2 (2018), 67–73.

[25] Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences* 11, 2 (2007), 77–83.

[26] Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. 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 (2002), 878.

[27] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 122–130.

[28] Iason Gabriel and Vafa Ghazavi. 2021. The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060* (2021).

[29] Ian Gemp, Kevin R. McKee, Richard Everett, Edgar A. Duéñez-Guzmán, Yoram Bachrach, David Balduzzi, and Andrea Tacchetti. 2020. D3C: Reducing the price of anarchy in multi-agent learning. In *Workshop on Cooperative AI at NeurIPS 2020*.

[30] Zohar Gilad, Ofra Amir, and Liat Levontin. 2021. 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*. 1–13.

[31] Donald W. Griesinger and James W. Livingston Jr. 1973. Toward a model of interpersonal motivation in experimental games. *Behavioral Science* 18, 3 (1973), 173–188.

[32] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[33] Clayton Hutto and Eric Gilbert. 2014. 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.

[34] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[35] Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47, 2 (1979), 263–292.

[36] Jurgis Karpus, Adrian Krüger, Julia Tovar Verba, Bahador Bahrami, and Ophelia Deroy. 2021. Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience* (2021), 102679.

[37] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David Newman, and Allison Woodruff. 2019. 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*. 627–637. https://doi.org/10.1145/3461702.3462605

[38] Nicolas Kervyn, Susan T. Fiske, and Chris Malone. 2012. Brands as intentional agents framework: How perceived intentions and ability can map brand perception. *Journal of Consumer Psychology* 22, 2 (2012), 166–176.

[39] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[40] Peter Kollock. 1998. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214.

[41] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 464–473.

[42] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* (2017).

[43] Wim B. G. Liebrand and Charles G. McClintock. 1988. 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 (1988), 217–230.

[44] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* (1932).

[45] Michael L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[46] Edward Lockhart, Neil Burch, Nolan Bard, Sebastian Borgeaud, Tom Eccles, Lucas Smaira, and Ray Smith. 2020. Human-agent cooperation in bridge bidding. *arXiv preprint arXiv:2011.14124* (2020).

[47] Kevin R. McKee, Xuechunzi Bai, and Susan T. Fiske. 2021. Understanding human impressions of artificial intelligence. *PsyArXiv* (2021). https://doi.org/10.31234/osf.io/5ursp

[48] Kevin R. McKee, Xuechunzi Bai, and Susan T. Fiske. 2022. Warmth and competence in human-agent cooperation. *arXiv preprint arXiv:2201.13448* (2022).

[49] Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duèñez-Guzmán, Edward Hughes, and Joel Z. Leibo. 2020. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 869–877.

[50] David Miller. 1983. Constraints on freedom. *Ethics* 94, 1 (1983), 66–86.

[51] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*. 1928–1937.

[52] Ryan O. Murphy and Kurt A. Ackermann. 2014. Social Value Orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review* 18, 1 (2014), 13–41.

[53] Nico J. D. Nagelkerke. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78, 3 (1991), 691–692.

[54] Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4, 2 (2013), 133–142.

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[56] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.
[57] Delroy L. Paulhus and Simine Vazire. 2007. The self-report method. In *Handbook of Research Methods in Personality Psychology*. Vol. 1. Guilford, 224–239.
[58] Eyal Pe'er, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.
[59] Eyal Pe'er, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2021), 1–20.
[60] Alexander Peysakhovich and Adam Lerer. 2018. Consequentialist conditional cooperation in social dilemmas with imperfect information. In *Proceedings of the 8th International Conference on Learning Representations*.
[61] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature* 597 (2021), 672–677.
[62] Glenn Reeder. 2009. Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological inquiry* 20, 1 (2009), 1–18.
[63] Byron Reeves and Jeff Hancock. 2020. Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior* 1, 1 (2020). https://doi.org/10.1037/tmb0000018
[64] Ann Marie T. Russell and Susan T. Fiske. 2008. It's all relative: Competition and status drive interpersonal perception. *European Journal of Social Psychology* 38, 7 (2008), 1193–1201.
[65] Paul A. Samuelson. 1948. Consumption theory in terms of revealed preference. *Economica* 15, 60 (1948), 243–253.
[66] Marcus M. Scheunemann, Raymond H. Cuijpers, and Christoph Salge. 2020. 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)*. IEEE, 1340–1347.
[67] Reinhard Selten. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4, 1 (1975), 25–55.
[68] Verónica Sevillano and Susan T. Fiske. 2016. Warmth and competence in animals. *Journal of Applied Social Psychology* 46, 5 (2016), 276–293.
[69] Claude E. Shannon. 1950. Programming a computer for playing chess. *Philos. Mag.* 41, 314 (1950), 256–275.
[70] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
[71] Satinder P. Singh, Andrew G. Barto, and Nuttapong Chentanez. 2005. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*.
[72] Ho Chit Siu, Jaime D. Pena, Kimberlee C. Chang, Edenna Chen, Yutai Zhou, Victor J. Lopez, Kyle Palko, and Ross E. Allen. 2021. Evaluation of human-AI Teams for learned and rule-based agents in Hanabi. *arXiv preprint arXiv:2107.07630* (2021).
[73] Robert Slonim and Ellen Garbarino. 2008. Increases in trust and altruism from partner selection: Experimental evidence. *Experimental Economics* 11, 2 (2008), 134–153.
[74] Brian Stanton and Theodore Jensen. 2021. Trust and artificial intelligence. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087
[75] Marilyn Strathern. 1997. 'Improving ratings': Audit in the British University system. *European Review* 5, 3 (1997), 305–321.
[76] DJ Strouse, Kevin R. McKee, Matthew Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*.
[77] Paul Tylkin, Goran Radanovic, and David C. Parkes. 2021. Learning robust helpful behaviors in two-player cooperative Atari environments. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 1686–1688.
[78] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.
[79] Jeffrey M. Zacks. 2004. Using movement and intentions to understand simple events. *Cognitive Science* 28, 6 (2004), 979–1008.