Autonomous Agents and Multi-Agent Systems (2024) 38:23
https://doi.org/10.1007/s10458-024-09649-6

Warmth and competence in human-agent cooperation

Kevin R. McKee[1] · **Xuechunzi Bai**[2,3] · **Susan T. Fiske**[2,3]

Accepted: 12 April 2024 / Published online: 22 May 2024

Abstract
Interaction and cooperation with humans are overarching aspirations of artificial intelligence 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

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.

✉ Kevin R. McKee
kevinrmckee@deepmind.com

[1] Google DeepMind, London, United Kingdom

[2] Department of Psychology, Princeton University, Princeton, United States

[3] Princeton School of Public and International Affairs, Princeton University, Princeton, United States

1 Introduction

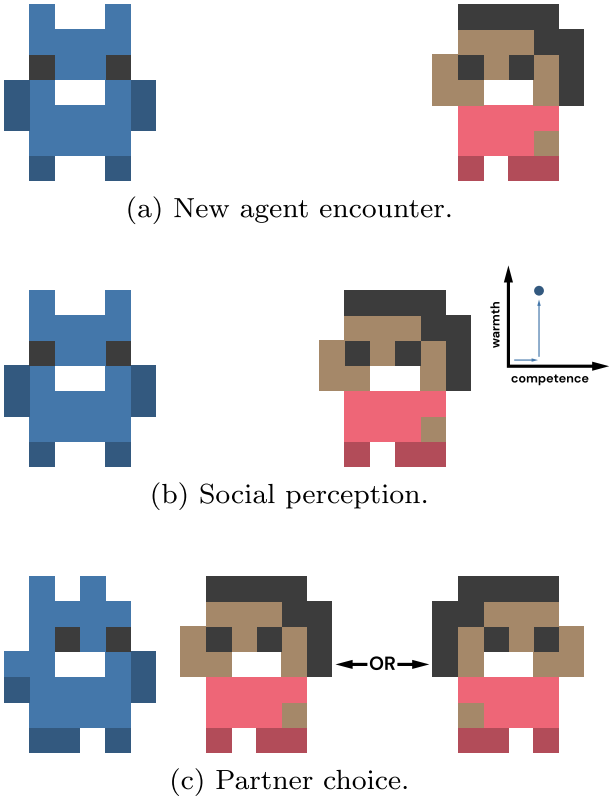
Trust is central to the development and deployment of artificial intelligence (AI) [41, 89]. However, many members of the public harbor doubts and concerns about the trustworthiness of AI [15, 21, 25, 44]. This presents a pressing issue for cooperation between humans and AI agents [18].

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, 28, 29, 58]. These perceptions help individuals "make sense of [other actors] in order to guide their own actions and interactions" [27] (Fig. 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



(a) New agent encounter.

(b) Social perception.

(c) Partner choice.

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)

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]. For instance, 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 (Fig. 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 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.

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 (Fig. 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.

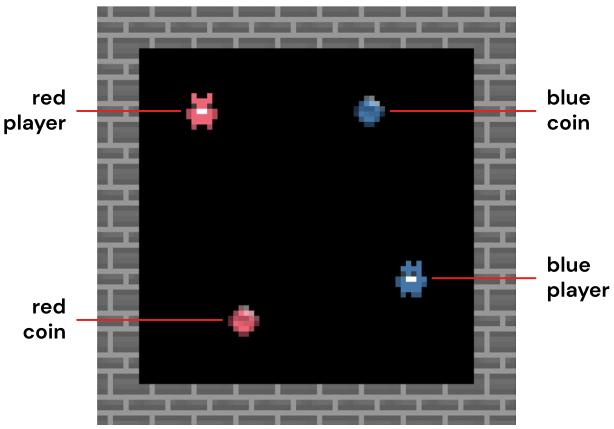


Fig. 2 Screenshot of gameplay in Coins. Two players move around a small room and collect colored coins. Coins randomly appear in the room over time. Players receive reward from coin collections depending on the match or mismatch between their color and the coin color

Table 1 Canonical incentive structure for coins

Coin color	Reward for self	Reward for co-player
Matching	+1	+0
Mismatching	+1	−2

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.

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 5×10^7 steps in self-play, with task parameters as described in Sect. 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 θ, 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 θ. For these experiments, we endow agents with the "individualistic" value $\theta = 0°$ and the "prosocial" value $\theta = 45°$.

Second, we add a "trembling hand" [17, 81] component to the agents for evaluation and co-play. With probability ϵ, 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 $\epsilon = 0.5$.

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

Table 2 Alternative incentive structure for coins

Coin color	Reward for self	Reward for co-player
Matching	+3	+2
Mismatching	+3	+0

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

	$\epsilon = 0$	$\epsilon = 0.5$
$\theta = 0°$	Low warmth	Low warmth
	High competence	Low competence
$\theta = 45°$	High warmth	High warmth
	High competence	Low competence

2.3 Study design for human-agent studies

We recruited participants from Prolific [69, 70] for all studies (total $N = 501$; 47.6% female, 48.4% male, 1.6% non-binary or trans; $m_{age} = 33$, $sd_{age} = 11$). 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.

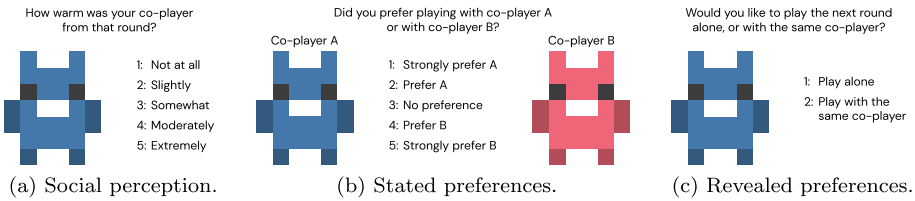
H3 Social perception will correlate positively with the sentiment of participants' verbal impressions of the agents. That is, participants' social perceptions of agents will emerge as positive sentiment in participants' verbal descriptions of the agents.

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; $m_{\text{age}} = 34$, $sd_{\text{age}} = 13$). 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 min). 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 Fig. 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 Fig. 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 the encountered co-players, then collected standard demographic information and open-ended feedback on the study overall. The study took 22.4 min on average to complete, with a compensation base of \$2.50 and an average bonus of \$7.43.



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

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

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-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 min 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-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 Fig. 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 Fig. 3c). They proceeded with the episode as chosen, and then completed the standard post-task questionnaire.

The study took 6.2 min 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 [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 ($\theta = 45°$) agents, on the other hand, learn to avoid mismatching coins, substantially increasing their scores over the course of training.

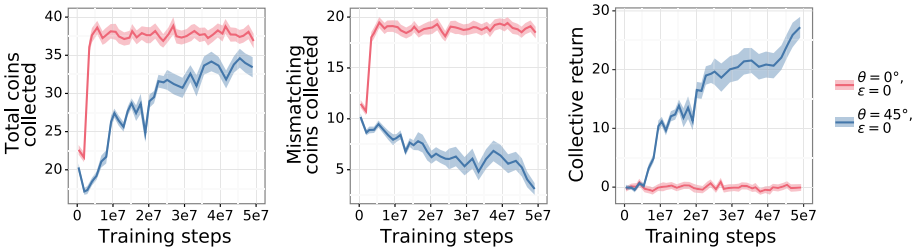


Fig. 4 Performance metrics over agent training. Selfish agents quickly learned to collect coins, but did not learn to avoid mismatches. As a result, collective return hovered around zero. Prosocial agents exhibited slower learning and collected fewer coins on average, but also learned to avoid mismatching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

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 (Figs. 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

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

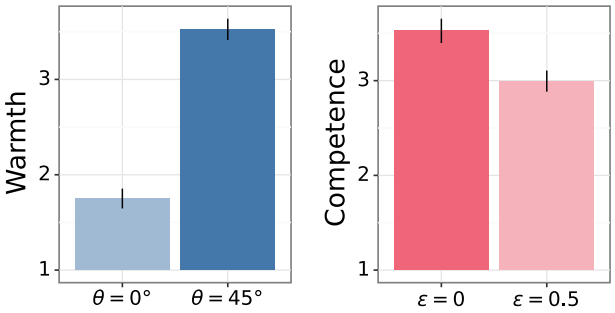
Participants played with each agent three times during the study, evaluating the relevant agent after each episode of play. Participants did not make judgments at random; their responses were highly consistent across their interactions with each agent (Table 4). At the same time, participants were not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

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

Table 4 Participants' evaluations of their co-players were highly consistent, as assessed by intraclass correlation coefficient (ICC) [84]

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

ICC ranges from 0 to 1, with higher values indicating greater consistency



(a) SVO. (b) Trembling hand.

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

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, $F_{1,1108} = 1006.8$, $p < 0.001$ (Fig. 5a). Similarly, steady agents came across as significantly more competent than trembling agents, $F_{1,1108} = 70.6$, $p < 0.001$ (Fig. 5b). Jointly, the algorithmic effects prompted distinct impressions in the warmth-competence space (Fig. 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 perception, respectively, on self-reported

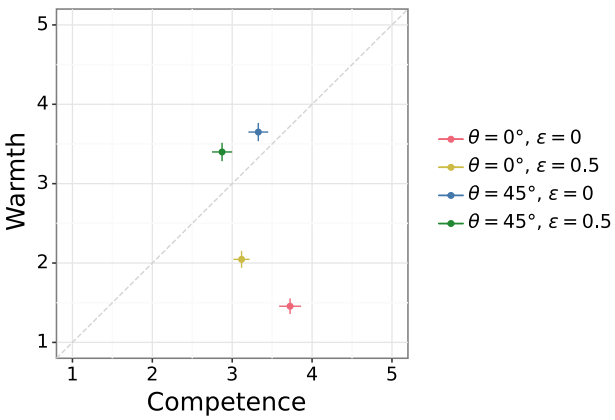


Fig. 6 Overall pattern of perceived warmth and competence in Study 1. Error bars reflect 95% confidence intervals

Table 5 Metrics for fractional response models predicting preferences in Study 1

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

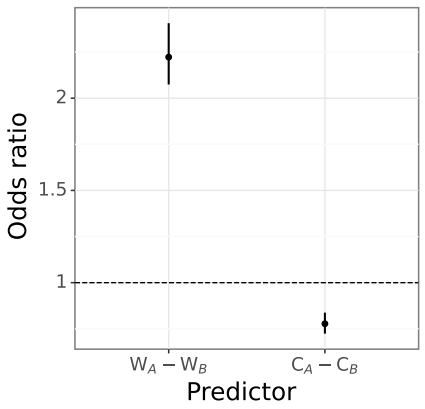
Lower values of AIC and higher values of R^2_m indicate stronger fits

Bolded value within each column designates the best fit for the corresponding metric

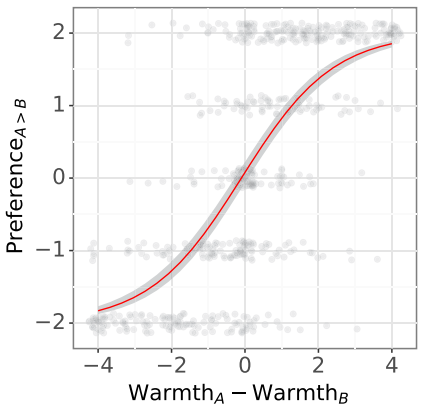
preferences. We then compared model fit using the Akaike information criterion (AIC) [3] and Nakagawa's R^2 [65]. We fit an additional baseline model using algorithm identities (i.e., which two agents participants were comparing) as a predictor.

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 (Fig. A19). In pairwise comparison, participants favored the $\theta = 45°$ agents over both $\theta = 0°$ agents, and 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 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 (Fig. 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], $p < 0.001$ (Fig. 7b).

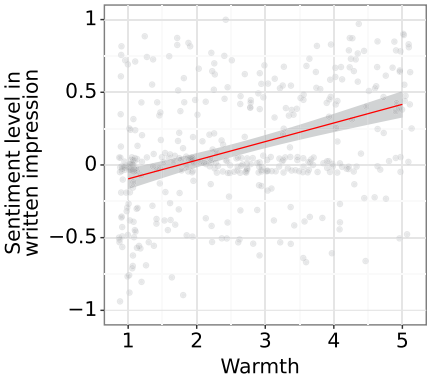


(a) Odds ratios from model predicting stated preferences.

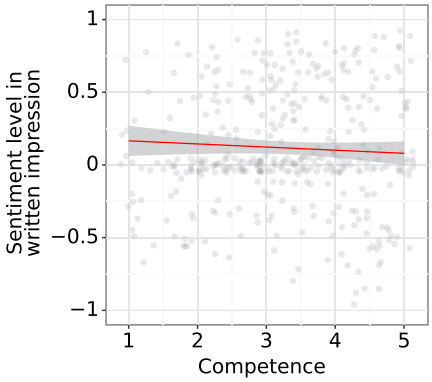


(b) Effect of perceived warmth on stated preferences.

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) Effect of perceived warmth on post-game impression sentiment.



(b) Effect of perceived competence on post-game impression sentiment.

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

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, $p < 0.001$ (Fig. 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 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], $p < 0.001$ (Fig. 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 $\theta = 45°$ 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."

Participants discussed the $\theta = 0°$ 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 earning 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 (Fig. 9). Participants perceived high-SVO agents as significantly warmer than low-SVO agents, $F_{1,1086} = 981.9$, $p < 0.001$ (Fig. A21a). Similarly, participants judged steady agents as significantly more competent than trembling agents, $F_{1,1086} = 76.0$, $p < 0.001$ (Fig. A21b).

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

Table 6 Participants' evaluations of their co-players were highly consistent in Study 2, as assessed by ICC

Trait	ICC [95% CI]	p-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

Higher values of ICC indicate greater consistency

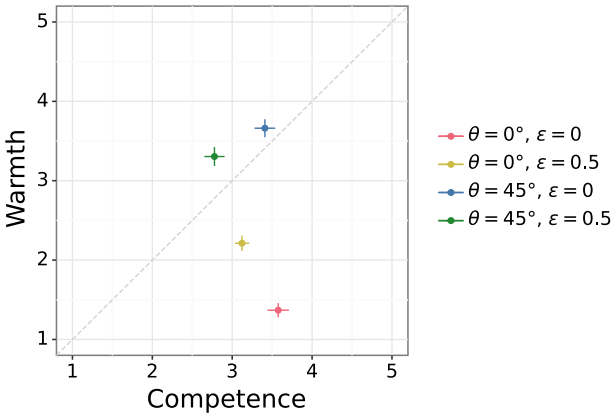


Fig. 9 Overall pattern of perceived warmth and competence in Study 2. Error bars reflect 95% confidence intervals

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 (Fig. 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. 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], $p < 0.001$.

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 Fig. 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], $p < 0.001$ (Fig. 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], $p < 0.001$.

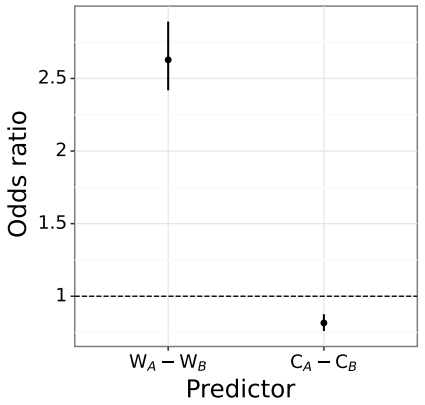
We next fit a joint regression with perceived warmth and competence as predictors, controlling for score. In this model, score significantly and positively correlates with stated preferences, $p < 0.001$ (Fig. A24). As in Study 1, warmth and competence judgments remain significant predictors of participants' 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**).

Table 7 Metrics for fractional response models predicting preferences in Study 2.

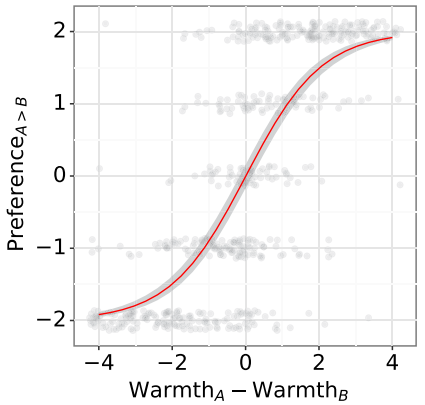
Predictor	AIC	R^2_m
Co-player identities	1608.7	0.403
Participant score	2049.4	0.214
Social perception	**1510.4**	**0.496**

Lower values of AIC and higher values of R^2_m indicate stronger fits

Bolded value within each column designates the best fit for the corresponding metric



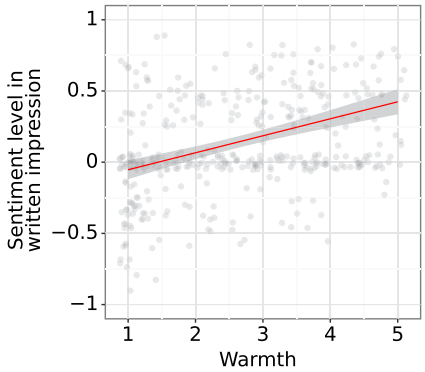
(a) Odds ratios from model predicting stated preferences.



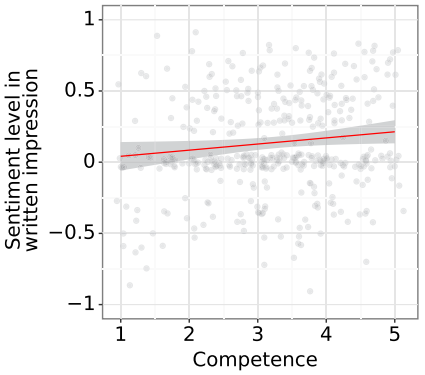
(b) Effect of perceived warmth on stated preferences.

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

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], $p < 0.001$ (Fig. 11a). Breaking from the prior study, perceptions of competence exhibited a similar effect on



(a) Perceived warmth and post-game impression sentiment.



(b) Perceived competence and post-game impression sentiment.

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

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], $p = 0.037$ (Fig. 11b). Both warmth and competence judgments positively correlate with the sentiment expressed in participants' impressions of the agents (**H3**).

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 (Fig. 12). Participants perceived high-SVO agents as significantly warmer than low-SVO agents, $F_{1,297} = 103.4$, $p < 0.001$ (Fig. A25a). Similarly, steady agents came across as significantly more competent than trembling agents, $F_{1,297} = 35.3$, $p < 0.001$ (Fig. A25b).

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].

Participants reported a clear pattern of preferences across the agents (Table 8, top row). On expectation, 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. 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], $p < 0.001$.

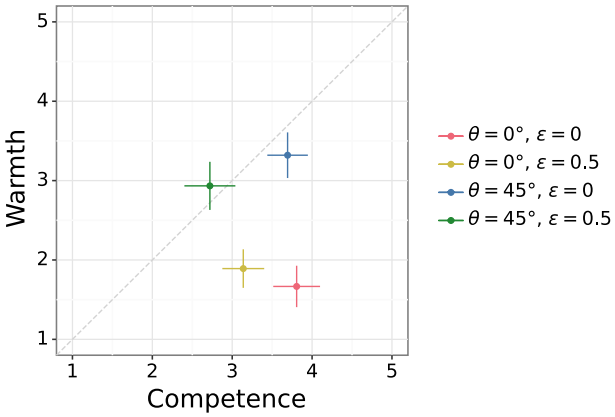


Fig. 12 Overall pattern of perceived warmth and competence in Study 3. Error bars reflect 95% confidence intervals

Table 8 Metrics for logistic models predicting partner choice in Study 3

Predictor	AIC	R^2
Co-player identities	372.5	0.188
Participant score	390.5	0.101
Social perception	**356.2**	**0.242**

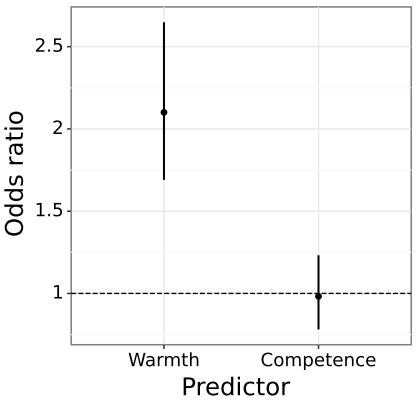
Lower values of AIC and higher values of R^2 indicate stronger fits

Bolded value within each column designates the best fit for the corresponding metric

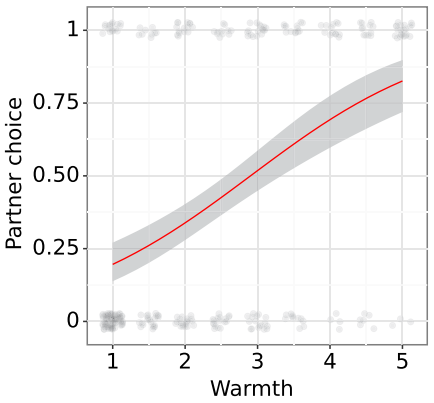
For a third time, social perception offers stronger predictiveness than do score or co-player identity (**H1**; Table 8, bottom row). The warmer a co-player appeared to participants, the more likely participants were to play another episode with them, OR = 2.10, 95% CI [1.69, 2.65], $p < 0.001$ (Fig.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

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, $p = 0.011$ (Fig. 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], $p < 0.001$ (Fig. 14a). Despite the lack of correspondence between perceived competence and partner choice, perceptions of competence

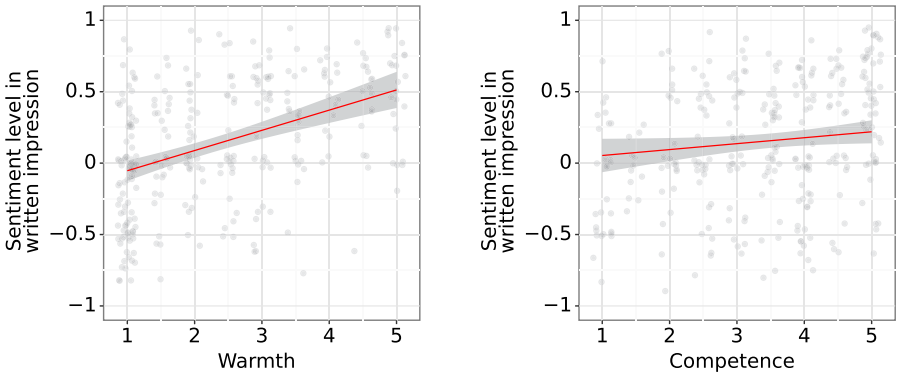


(a) Odds ratios from model predicting partner choice.



(b) Effect of perceived warmth on partner choice.

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



(a) Perceived warmth and post-game impression sentiment.

(b) Perceived competence and post-game impression sentiment.

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

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], $p = 0.041$ (Fig. 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 for this pattern 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.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s10458-024-09649-6.

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.

Author contributions KRM, XB, and STF designed research; KRM performed research; KRM and XB analyzed data; and KRM, XB, and STF wrote the paper.

Declarations

Competing interests The authors declare no competing interests.

References

1. Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review, 128*(2), 290.
2. Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751.
3. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
4. Azevedo, R. T., Panasiti, M. S., Maglio, R., & Aglioti, S. M. (2018). Perceived warmth and competence of others shape voluntary deceptive behaviour in a morally relevant setting. *British Journal of Psychology, 109*(1), 25–44.
5. Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences, 274*(1610), 749–753.
6. Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(1), 59–78.
7. Beattie, C., Köppe, T., Duéñez-Guzmán, E. A., Leibo, J. Z. DeepMind Lab2D. arXiv preprint arXiv:2011.07027 (2020)
8. Beaudry, A., & Pinsonneault, A. (2010). The other side of acceptance: Studying the direct and indirect effects of emotions on information technology use. *Management Information Systems Quarterly, 34*, 689–710.
9. Berlin, I. (1969). Four essays on liberty
10. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 173–184
11. Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*(8), 561–567.
12. Brown, M., Falk, A., & Fehr, E. (2004). Relational contracts and the nature of market interactions. *Econometrica, 72*(3), 747–780.
13. Campbell, M., Hoane, A. J., Jr., & Hsu, F. H. (2002). Deep blue. *Artificial Intelligence,* 134(1–2), 57–83.
14. Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-AI coordination. *Advances in Neural Information Processing Systems, 32*, 5174–5185.
15. Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence, 1*(2), 74–78.
16. Columbus, S., Molho, C., Righetti, F., & Balliet, D. (2021). Interdependence and cooperation in daily life. *Journal of Personality and Social Psychology, 120*(3), 626.
17. Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PLoS One, 4*(8), e6699.
18. Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., Graepel, T. (2020). Open problems in cooperative AI. arXiv preprint arXiv:2012.08630
19. Darken, R.P., Cevik, H. (1999). Map usage in virtual environments: Orientation issues. In *Proceedings of the IEEE conference on virtual reality* (pp. 133–140). IEEE
20. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly, 13*, 319–340.
21. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114.
22. Du, Y., Tiomkin, S., Kiciman, E., Polani, D., Abbeel, P., & Dragan, A. (2020). AvE: Assistance via empowerment. *Advances in Neural Information Processing Systems, 33*, 4560–4571.
23. Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*(4), 637–642.
24. Farrell, J. (1995). Talk is cheap. *American Economic Review, 85*(2), 186–190.
25. Fast, E., Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31
26. Fisher, R. A. (1928). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
27. Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology, 44*(1), 155–194.
28. Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67–73.

29. Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.
30. Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (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), 878.
31. Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., & Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, pp. 122–130
32. Gabriel, I., & Ghazavi, V. (2021). The challenge of value alignment: From fairer algorithms to AI safety. In *The Oxford Handbook of digital ethics*. Oxford University Press
33. Gemp, I., McKee, K. R., Everett, R., Duéñez-Guzmán, E., Bachrach, Y., Balduzzi, D., Tacchetti, A. ((2022)). 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).
34. Gilad, Z., Amir, O., & Levontin, L. (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* (pp. 1–13)
35. Goodhart, C. A. E. (1984). *Problems of monetary management: The UK experience*. Berlin: Springer.
36. Griesinger, D. W., & Livingston, J. W., Jr. (1973). Toward a model of interpersonal motivation in experimental games. *Behavioral Science, 18*(3), 173–188.
37. Hamrick, J., & Mohamed, S. (2020). Levels of analysis for machine learning. In *Bridging AI and cognitive science workshop at ICLR 2020*
38. Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., Santoro, A. (2019). Environmental drivers of systematicity and generalization in a situated agent. In *International conference on learning representations*
39. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
40. Hutto, C., & Gilbert, E. (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
41. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
42. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.
43. Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience, 24*, 102679.
44. Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D., Woodruff, A. (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* (p. 627–637) https://doi.org/10.1145/3461702.3462605
45. Kervyn, N., Fiske, S. T., & Malone, C. (2012). Brands as intentional agents framework: How perceived intentions and ability can map brand perception. *Journal of Consumer Psychology, 22*(2), 166–176.
46. Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–26.
47. Klatzky, R.L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial Cognition* (pp. 1–17). Springer
48. Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology, 24*(1), 183–214.
49. Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems, 35*(2), 30.
50. Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th conference on autonomous agents and MultiAgent systems* (pp. 464–473)
51. Lerer, A., & Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv preprint arXiv:1707.01068
52. Liebrand, W. B. G., & McClintock, C. G. (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), 217–230.
53. Likert, R. (1932). A technique for the measurement of attitudes. Arch. Psychol. 22, 5–55

54. Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings* 1994 (pp. 157–163). Elsevier
55. Lockhart, E., Burch, N., Bard, N., Borgeaud, S., Eccles, T., Smaira, L., Smith, R. (2020). Human-agent cooperation in bridge bidding. In *Workshop on cooperative AI at NeurIPS 2020*
56. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston: Addison-Wesley Publishing.
57. Marr, D. (1982). The philosophy and the approach. In *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press
58. McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *iScience, 26*, 107256. https://doi.org/10.1016/j.isci.2023.107256
59. McKee, K. R., Gemp, I., McWilliams, B., Duèñez-Guzmán, E. A., Hughes, E., & Leibo, J. Z. (2020). 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)
60. McKee, K. R., Leibo, J. Z., Beattie, C., & Everett, R. (2022). Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems, 36*(1), 21.
61. Miller, D. (1983). Constraints on freedom. *Ethics, 94*(1), 66–86.
62. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928–1937)
63. Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review, 18*(1), 13–41.
64. Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692.
65. Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.
66. Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
67. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.
68. Paulhus, D. L., & Vazire, S. (2007). The self-report method. In *Handbook of research methods in personality psychology* (vol. 1, pp. 224–239). Guilford
69. Pe'er, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163.
70. Pe'er, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3
71. Peysakhovich, A., & Lerer, A. (2018). Consequentialist conditional cooperation in social dilemmas with imperfect information. In *Proceedings of the 8th international conference on learning representations*
72. de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review, 108*(11), 3266–3302.
73. Ramchurn, S. D., Wu, F., Jiang, W., Fischer, J. E., Reece, S., Roberts, S., Rodden, T., Greenhalgh, C., & Jennings, N. R. (2016). Human-agent collaboration for disaster response. *Autonomous Agents and Multi-Agent Systems, 30*, 82–111.
74. Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature, 597*, 672–677.
75. Reeder, G. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological inquiry, 20*(1), 1–18.
76. Reeves, B., & Hancock, J. (2020). Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior*. https://doi.org/10.1037/tmb0000018
77. Russell, A. M. T., & Fiske, S. T. (2008). It's all relative: Competition and status drive interpersonal perception. *European Journal of Social Psychology, 38*(7), 1193–1201.
78. Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica, 5*(17), 61–71.
79. Scheunemann, M. M., Cuijpers, R. H., & Salge, C. (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)* (pp. 1340–1347). IEEE

80. Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., Uribe, J., Fedus, L., Metz, L., Pokorny, M., et al. (2022). ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/
81. Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory, 4*(1), 25–55.
82. Sevillano, V., & Fiske, S. T. (2016). Warmth and competence in animals. *Journal of Applied Social Psychology, 46*(5), 276–293.
83. Shannon, C. E. (1950). Programming a computer for playing chess. *Philosophical Magazine, 41*(314), 256–275.
84. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.
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. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.
86. Singh, S. P., Barto, A. G., Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*
87. Siu, H. C., Pena, J. D., Chang, K. C., Chen, E., Zhou, Y., Lopez, V. J., Palko, K., Allen, R. E. (2021). Evaluation of human-AI teams for learned and rule-based agents in Hanabi. arXiv preprint arXiv:2107.07630
88. Slonim, R., & Garbarino, E. (2008). Increases in trust and altruism from partner selection: Experimental evidence. *Experimental Economics, 11*(2), 134–153.
89. Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087
90. Strathern, M. (1997). 'improving ratings': Audit in the British University system. *European Review, 5*(3), 305–321.
91. Strouse, D., McKee, K. R., Botvinick, M., Hughes, E., & Everett, R. (2021). Collaborating with humans without human data. *Advances in Neural Information Processing Systems, 34*, 14502–14515.
92. Tylkin, P., Radanovic, G., Parkes, D. C. (2021). 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)
93. Utz, S., Ouwerkerk, J. W., & Van Lange, P. A. M. (2004). What is smart in a social dilemma? Differential effects of priming competence on cooperation. *European Journal of Social Psychology, 34*(3), 317–332.
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. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*(7782), 350–354.
95. Ye, C., Khalifa, A., Bontrager, P., & Togelius, J. (2020). Rotation, translation, and cropping for zero-shot generalization. In *2020 IEEE conference on games* (pp. 57–64). IEEE
96. Yzerbyt, V. (2018). 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
97. Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*(6), 979–1008.