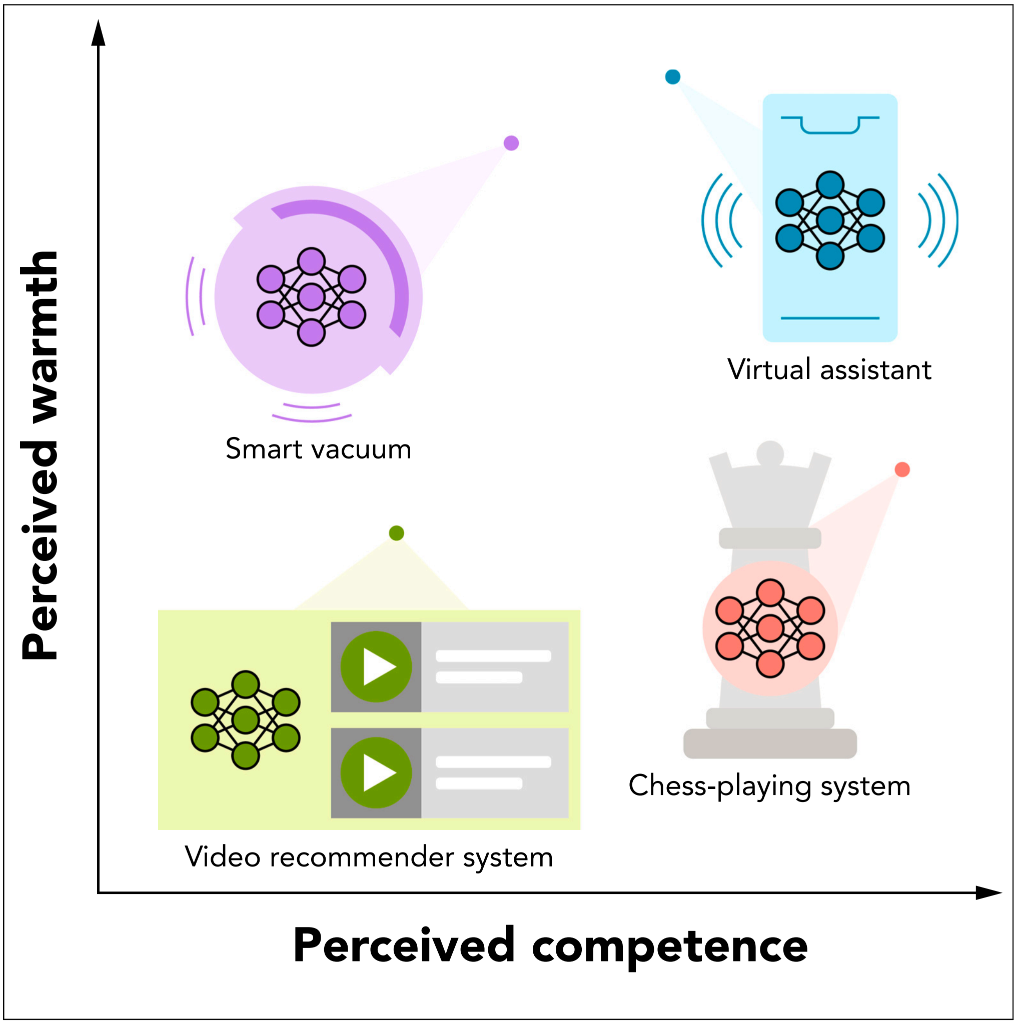
iScience

CellPress OPEN ACCESS

Article

# Humans perceive warmth and competence in artificial intelligence



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

kevinrmckee@deepmind.com

## Highlights

Nine studies show that humans think of A.I. in social terms

Modern A.I. systems evoke perceptions of both warmth and competence

People perceive systems that pursue interests aligned with human interests as warmer

People see systems that operate independently from human oversight as more competent

McKee et al., iScience *26*, 107256
August 18, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.107256

iScience

CellPress
OPEN ACCESS

Article

# Humans perceive warmth and competence in artificial intelligence

Kevin R. McKee,[1,4,*] Xuechunzi Bai,[2,3] and Susan T. Fiske[2,3]

## SUMMARY

**Artificial intelligence (A.I.) increasingly suffuses everyday life. However, people are frequently reluctant to interact with A.I. systems. This challenges both the deployment of beneficial A.I. technology and the development of deep learning systems that depend on humans for oversight, direction, and regulation. Nine studies ($N$ = 3,300) demonstrate that social-cognitive processes guide human interactions across a diverse range of real-world A.I. systems. Across studies, perceived warmth and competence emerge prominently in participants' impressions of A.I. systems. Judgments of warmth and competence systematically depend on human-A.I. interdependence and autonomy. In particular, participants perceive systems that optimize interests aligned with human interests as warmer and systems that operate independently from human direction as more competent. Finally, a prisoner's dilemma game shows that warmth and competence judgments predict participants' willingness to cooperate with a deep-learning system. These results underscore the generality of intent detection to perceptions of a broad array of algorithmic actors.**

## INTRODUCTION

Artificial intelligence (A.I.) systems and machine learning algorithms play a central role in everyday life. People receive suggestions from recommender systems when listening to music,[1] watching videos or movies,[2,3] and browsing online content.[4] Individuals rely on the speech recognition capabilities of virtual assistants to schedule their days and help manage their chores.[5] Game experts and enthusiasts compete in matches against agents trained with reinforcement learning.[6–8] In certain parts of the world, drivers share the road with autonomous vehicles.[9] Proposals for long-term applications of A.I. are sweeping, with roles for A.I. in domains such as education and health care,[10] across both developing and developed economies.[11]

While the integration of A.I. into society heralds many potential benefits (see Christakis[12]), it also raises critical issues surrounding risks, trust, and public sentiment.[13,14] A substantial proportion of the public fears that A.I. systems will have negative effects on their lives,[15,16] and people are often reluctant to personally adopt A.I. systems.[17,18] These attitudes have mixed consequences; they impede the adoption of beneficial A.I. systems by everyday individuals and users, and simultaneously offer a measure of protection against malicious uses of A.I. In parallel, they challenge the development of novel A.I. technologies that depend on humans as sources of direction, instruction, auditing, oversight, reward, and data.[19–22]

To facilitate positive relationships between human and A.I., technologists and social scientists must understand human impressions of A.I. technology. One social-cognitive process by which people form impressions of others—and thus control uncertainty in their social interactions—is intent detection.[23–26] Non-human actors, including robots, can elicit such judgments.[27,28] However, relatively few empirical studies have investigated the evaluations prompted by A.I. (though see Ashktorab et al.[29] and Khadpe et al.[30]).

Here, tools from social cognition research help to evaluate impressions of a broad, diverse range of A.I. systems and to test potential antecedents of these impressions. Intent detection proves central to impressions of A.I. Three initial studies assess the foundations of intent detection, posing the question: do people perceive A.I. systems as social actors, as opposed to mere tools? Three subsequent studies test whether warmth and competence, two important dimensions in the social perception of humans, are present in

[1] DeepMind, N1C 4DN London, UK

[2] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

[3] School of Public and International Affairs, Princeton University, Princeton, NJ 08540, USA

[4] Lead contact

*Correspondence: kevinrmckee@deepmind.com

https://doi.org/10.1016/j.isci.2023.107256

impressions of A.I. systems. Do warmth and competence emerge in any consistent pattern in people's perceptions of A.I.? The next two studies explore factors that shape people's judgments of the warmth and competence of A.I. systems: how do perceived covariation of interests, perceived status, and perceived autonomy affect impressions of A.I. systems? In the final experiment, humans play economic games with reinforcement learning agents to investigate impressions of A.I. in incentive-compatible interactions. Do people's perceptions of warmth and competence influence their choices when they interact with A.I. systems?

## Artificial intelligence as social actor

At its core, an A.I. is a human-made process or system that makes decisions or solves problems.[31] Common problems addressed by A.I. systems include voice recognition (e.g., virtual assistants), preference prediction (e.g., recommender systems), and move selection in games (e.g., game-competitor systems). In recent years, technical advances (including the advent of deep learning methods[32]) have transformed modern A.I. and greatly expanded its capabilities. A.I. systems now pervade political and economic processes throughout society, as well as everyday personal and interpersonal interactions.[33] As a result, such systems represent a new class of actors meriting psychological study. Do any consistent patterns structure human impressions of A.I. systems? What factors and antecedents guide impression formation? How might these impressions influence human interactions with A.I. systems? Understanding these questions can reveal how humans navigate a world populated by algorithmic actors and guide policymakers as they deploy and regulate new systems.

Research on related categories of actors and agents provides lessons for new work on A.I. Of particular note are robots, machines that can operate with some degree of autonomy in physical environments.[34] Whereas decision-making and intelligence are central to A.I., physical embodiment defines robots.[35] (The two classes overlap: some robots include algorithms to guide their perception or actions, and thus robots sometimes qualify as A.I. systems.) Tellingly, people respond to robots as social actors rather than asocial objects.[36–38] Under certain conditions, individuals ascribe minds—and consequently, moral standing—to robots.[25,28,39,40]

Reflecting the physical embodiment defining robotics, this line of research often focuses on the effects of appearance (e.g., Reeves, Hancock, & Liu[41]), frequently for anthropomorphic robots (e.g., DiSalvo et al.[42] and Goetz, Kiesler, & Powers[43]). In contrast, the centrality of decision-making to A.I. raises particular concerns that often are not shared by roboticists.[35] For example, a prominent challenge particular to A.I. research concerns the fairness exhibited by machine learning algorithms as they interact with different communities (''algorithmic fairness'').[22,44] Similarly, a key design choice when training A.I. systems through reinforcement learning is the selection of goals and rewards to guide learning (''reward specification'').[45] The distinctions between A.I. and robots emphasize the value of understanding people's reactions to the A.I. systems they increasingly encounter.

## Warmth and competence: Fundamental dimensions of social perception

The relationship between human and A.I. varies widely across extant and proposed applications of A.I. technology. As noted, prominent A.I. systems interact with humans as competitors, assistants, and recommenders. In the future, people may engage with A.I. as resident to city planner, student to teacher, or patient to caretaker. Each of these roles suggests a different *structure of interdependence*[46] between human and A.I. interactants. Human impressions of an A.I. system will likely be shaped not only by the system's behavior but also by the relational context in which interaction unfolds. Relationships determine the other's perceived intent.

The Stereotype Content Model (SCM), developed in social cognition research, theorizes that the structure of interdependence is a critical determinant of social perception.[23,47] Work on the SCM has identified two primary dimensions of social perception: warmth (trustworthiness and friendliness) and competence (capability and confidence). Warmth and competence appear fundamental to impression formation, characterizing perception of other humans,[48] as well as impressions of non-human agents that appear to have intent. Such entities include animals,[49] consumer brands,[50] and robots.[51] In prior work, Khadpe et al.[30] demonstrated the relevance of these two dimensions to chatbots, one particular variety of A.I.—raising the possibility that perceptions of warmth and competence are fundamental to A.I. systems as a broader class.

The SCM theorizes that judgments of warmth are driven in particular by the *covariation of interests* between the perceiver and the perceived social actor.[47] The covariation of interests in an interaction is the degree to which partners' intents or outcomes correspond.[52] For example, two people whose interests align are more likely to perceive each other as warm; a pair with opposed motives will see each other as cold. Modern A.I. research grapples with analogous constructs. Developers may design A.I. systems with goals that are more or less supportive of human goals and interests. In reinforcement learning research, for example, a key question is how aligned rewards should be between human and A.I. interactants.[53] Expanding existing evidence[30] to a diverse range of A.I. domains, how warm will participants perceive algorithmic agents to be? We hypothesize that the roles and goal alignment of A.I. interactants will systematically affect judgments of their warmth.

Judgments of competence follow two hypotheses. First, research on the SCM has demonstrated that in human groups, social status reliably predicts judgments of competence.[23,54] People perceive individuals occupying high-status positions as more competent than those in lower-status roles. This pattern may reflect a tendency for people to automatically attribute status and success to a (dispositional) capability to deliver on one's intentions (see Ross[55]). Such attributions help to legitimate existing status hierarchies, providing a sense of certainty, stability, and fairness to those in both dominant and subordinate positions.[56] If the social evaluation of A.I. agents recruits the same cognitive processes as the social evaluation of humans, this pattern could extend to A.I. as well. Higher-status A.I. systems may garner stronger attributions of competence. The definition of A.I. status remains to be specified.

Second, the study of A.I. often prompts discussions about the nature of autonomy and agency.[57–59] Pragmatically, developers can design A.I. systems to operate largely autonomously or to depend more closely on human interactants. These points suggest an alternative predictor for evaluations of an A.I. system's competence: its perceived autonomy. A.I. systems that rely heavily on human direction or that can be understood as simple input-output mappings might appear more like devices than intelligent agents (see also Dennett[60]). In contrast, A.I. systems acting according to complex operations beyond input-output mappings and with greater degrees of self-initiative might be readily perceived as especially competent. We hypothesize that attributions of autonomy drive judgments of competence.

## RESULTS

We investigated human impressions of A.I. systems in nine studies, recruiting a total of 3,300 human participants through the online platforms Mechanical Turk and Prolific.

### Studies 1, 2, and 3: A.I. systems as social actors

Three studies tested the initial premise that people perceive A.I. systems as social actors rather than asocial objects. Intentionality and goal-directed behavior distinguish actors from objects and tools.[60–62] Study 1 ($N$ = 30) therefore assessed whether participants attribute intentionality to several real-world A.I. systems, including a virtual assistant, a recommender system, and a game-competitor system. Participants judged to what extent each of the A.I. systems possessed intentions, goals, and "a mind of its own", using a 5-point scale. As an asocial baseline, participants also evaluated everyday tools with similar uses as the A.I. systems. Participants ascribed significantly more intentionality to A.I. systems than to tools with similar uses, $F(1, 173) = 22.7$, $p < 0.001$, $\omega_G^2 = 0.11$ (Figure 1A; see also Figure S1). Participants also perceived a mind in A.I. systems to a significantly greater degree than in tools, $F(1, 173) = 31.8$, $p < 0.001$, $\omega_G^2 = 0.15$ (Figure 1B; see also Figure S2).

Two subsequent studies evaluated whether participants apply social norms (in particular, the norm of politeness) to interactions with A.I. systems, a characteristic hallmark of social actors.[63–65] In Study 2 ($N$ = 30), participants reported the likelihood that they would feel gratitude and follow politeness norms, using a 7-point scale, after interactions with the same A.I. systems and tools as in Study 1. A.I. systems and inanimate tools elicited similar levels of gratitude for contributing toward participants' goals, $F(1, 173) = 0.03$, $p = 0.86$, $\omega_G^2 = 0.00$ (Figure S3). Nonetheless, participants reported a significantly higher behavioral intention to follow politeness norms when interacting with A.I. systems than with tools, $F(1, 173) = 4.7$, $p = 0.032$, $\omega_G^2 = 0.02$ (Figure 1C; see also Figure S4). Study 3 ($N$ = 30) replicated these results by assessing participants' reactions to a third party's interactions with A.I. Participants reported the appropriateness of third-party use of politeness norms toward the A.I. systems and tools from Studies 1 and 2, using a 7-point scale. Echoing the results from Study 2, participants endorsed the third-party use of politeness norms with

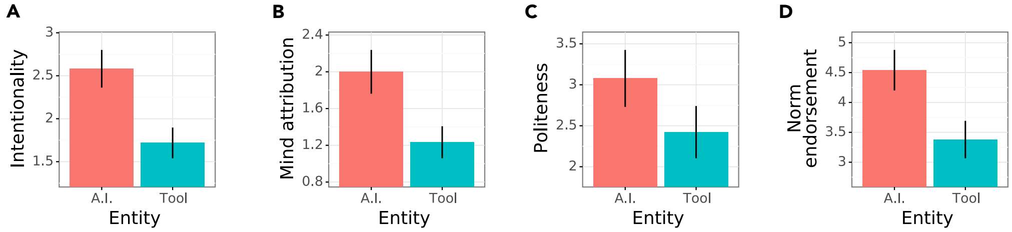


**Figure 1. People perceive A.I. systems as social actors**

Error bars reflect 95% confidence intervals.

(A) People attribute significantly greater intentionality to A.I. systems than to tools with similar uses. See also Figure S1.

(B) People believe that A.I. systems possess a mind of their own to a significantly greater degree than do tools with similar uses. See also Figure S2.

(C) Despite feeling similar levels of gratitude toward A.I. systems and tools for their use, people report being significantly more likely to follow politeness norms when interacting with A.I. systems. See also Figures S3 and S4.

(D) People endorse others' application of politeness norms to A.I. systems to a significantly greater degree than to tools. See also Figure S5.

A.I. systems to a significantly greater degree than with tools, $F(1, 167) = 21.8$, $p < 0.001$, $\omega_G^2 = 0.09$ (Figure 1D; see also Figure S5).

Altogether, these studies confirm that people perceive A.I. systems as more than mere tools, possessing intentionality and meriting the application of social norms. Given their standing as social actors, what sorts of impressions do A.I. systems prompt?

## Studies 4 and 5: Impressions of A.I. in natural language

Study 4 ($N = 99$) tested the spontaneous emergence of perceived warmth and competence in impressions of prominent examples of A.I. technology. The study gathered naturalistic content of impressions through written free-responses about these A.I. systems, instructing participants to write several sentences about their impression of each A.I. system.

The study provided participants with three different A.I. system roles to discuss: game competitors (e.g., AlphaGo), virtual assistants (e.g., Siri), and recommender systems (e.g., the movie recommendation system used by Netflix). Participants additionally evaluated several examples of A.I. technology falling outside these roles, given their prominence in public or scholarly discourse. These instructive examples included Roomba (a "social robot"),[66,67] self-driving cars,[68] and drones.[69,70]

An automated dictionary tool evaluated the content of participant responses, analyzing each response along a number of perceptual dimensions.[71] The tool coded each token in a response as a categorical indicator (positively valenced content, negatively valenced content, or absent) along each dimension, using a dictionary specific to that dimension. This analysis computed a "response coverage" variable for each written response, representing the proportion of tokens in the response that related to a given content dimension (regardless of negative or positive valence). Following the theoretical framework described in Abele et al.,[72] the analysis computed warmth through the simple combination of the morality and sociability dictionaries, and competence through the simple combination of the ability and assertiveness dictionaries.

The free-response data collected in Study 4 demonstrate the importance of warmth and competence in human impressions of real-world A.I. systems. On average, 9.4% of the content of participants' impressions related directly to warmth, and 10.7% related to competence (Figure 2A). In theoretical terms, warmth has two facets: perceived morality and sociability.[72] Similarly, competence comprises perceived ability and assertiveness. These perceptual subdimensions accounted for a particularly large proportion of impression content relative to other common perceptual dimensions (Figure 2B and Tables S1–S3). Participants frequently discussed systems in terms of ability (e.g., describing individual systems as "quite reliable most of the time for most basic tasks" or "the most capable"), morality ("somewhat rudimentary and untrustworthy"), and sociability ("the most bland out of other assistants"). Impressions also contained a relatively large amount of occupation-related information, potentially reflecting concerns surrounding A.I. and labor displacement.[13,14,73] Though both warmth and competence perfused participants'

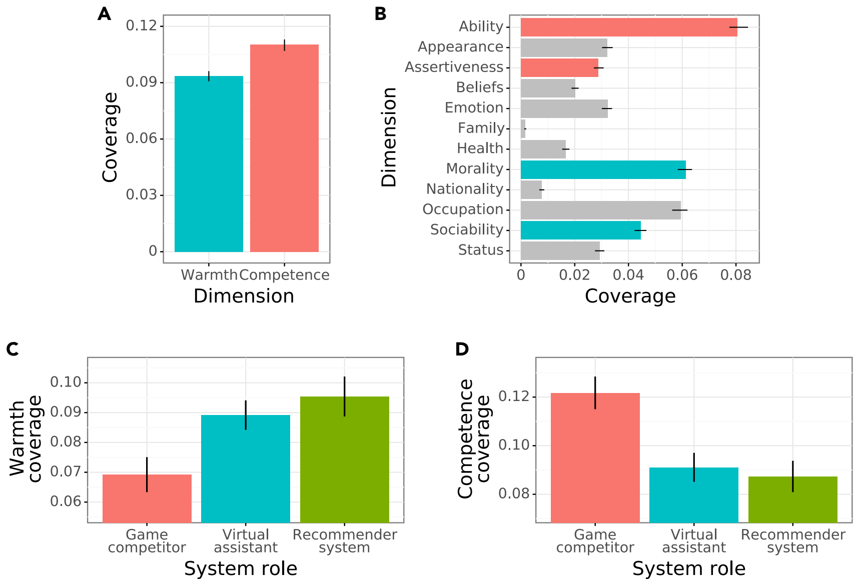


**Figure 2. Warmth and competence emerge as prominent dimensions in impressions of real-world A.I. systems**

Error bars indicate 95% confidence intervals.

(A) On average, impressions of the A.I. systems contained significantly more competence-related content than warmth-related content. See also Figure S6.

(B) Warmth and competence-related content appears in impressions at significantly higher levels relative to other common perceptual dimensions. Morality and sociability content constitutes the warmth dimension; ability and assertiveness content compose the competence dimension. See also Tables S1–S3 and Figure S7.

(C) An A.I. system's role significantly predicted warmth coverage. See also Table S4.

(D) Similarly, an A.I. system's role significantly predicted competence coverage. See also Table S5.

perceptions, competence-related content significantly exceeded warmth-related content, with an average marginal ratio of 1.14, 95% CI [1.09, 1.20], $p < 0.001$. On the whole, participants discussed their impressions through the language of warmth and competence—highlighting perceptions of ability and morality more than or as much as any other dimensions.

Systematic differences in warmth- and competence-related coverage emerge when the A.I. systems are categorized by system role (Figures 2C and 2D). Normalizing for response length, recommender systems ($m$ = 0.10, $sd$ = 0.06) and virtual assistants ($m$ = 0.09, $sd$ = 0.05) elicited the highest warmth coverage, followed by game-competitor systems ($m$ = 0.07, $sd$ = 0.05). A repeated-measures ANOVA confirmed that system role is a statistically significant predictor for warmth coverage, $F(2, 975) = 23.4$, $p < 0.001$, $\omega^2_G = 0.04$ (Figure 2C; see also Table S4). Competence coverage tended to be highest for game-competitor A.I. ($m$ = 0.12, $sd$ = 0.05), followed by virtual assistants ($m$ = 0.09, $sd$ = 0.05) and recommender systems ($m$ = 0.09, $sd$ = 0.06). A repeated-measures ANOVA indicated that system role significantly affected competence coverage, $F(2, 975) = 38.8$, $p < 0.001$, $\omega^2_G = 0.07$ (Figure 2D; see also Table S5).

In addition to the personal and game competition contexts examined in Study 4, a growing number of proposals suggest using A.I. systems in community-facing settings such as health care and hiring. These applications raise concerns about bias and fairness in algorithmic decision-making,[74–77] particularly for marginalized communities affected by the systems.[44] Study 5 ($N$ = 113) recruited participants to test whether warmth and competence emerge in impressions of systems in these ethically contested domains. The study provided participants with four application areas to discuss: hiring, health care, education, and facial recognition. As in Study 4, participants were instructed to write several sentences about their impression of each A.I. system. The same automated dictionary tool evaluated the content of participant responses, calculating coverage of responses by each of a number of perceptual dimensions.

The free-response data from Study 5 indicate that the prominence of warmth and competence carries over from competitions and personal domains to these ethically contested community-facing applications of

A.I. Warmth and competence content constituted 8.5% and 11.5% of the average impression, respectively (Figure S6). Morality and ability content remain particularly salient relative to other perceptual dimensions (Figure S7 and Tables S6–S8). Participants repeatedly foreground system competence (e.g., noting some systems "would outperform most any human" and wondering "how accurate or precise" others are), while also emphasizing the fairness and warmth that A.I. systems exhibit toward humans (e.g., assessing systems as "fair and impartial" or "impersonal and cold"). As in Study 4, impressions contained significantly more competence content than warmth content, with an average marginal ratio of 1.36, 95% CI [1.25, 1.47], $p < 0.001$. Taken together, Studies 4 and 5 illustrate that across both personal and community-facing contexts, individuals make sense of A.I. systems in terms of warmth and competence.

## Study 6: Impressions of A.I. and potential antecedents

In order to replicate the effects of system role on perceived warmth and competence and to explore potential mechanisms producing those effects, a sixth study shifted from gathering unconstrained written impressions of A.I. systems to collecting numeric judgments of their attributes along certain predetermined dimensions. Study 6 ($N = 154$) prompted participants with the same systems as in Studies 4 and 5, and asked them to evaluate each system's attributes. For each system, participants reported the warmth, competence, covariation of interests, status, and autonomy they perceived on a 5-point scale (see STAR Methods). Broadly, judgments of the A.I. systems in Study 6 echo the patterns previously observed in free-text impressions (Figure 3). Participants perceived A.I. systems as more competent than warm on average, $t(2, 279) = 34.5$, $p < 0.001$, $d = 0.72$. The systems fell mostly in the high-competence/low-warmth quadrant (Figure S8).

As before, game-competitor systems appeared high on competence ($m = 3.65$, $sd = 0.86$) but low on warmth ($m = 2.66$, $sd = 0.66$). Virtual assistants were high on both warmth ($m = 3.11$, $sd = 0.73$) and competence ($m = 3.68$, $sd = 0.72$). Finally, recommender systems came across as somewhat low on both warmth ($m = 2.70$, $sd = 0.83$) and competence ($m = 3.27$, $sd = 0.91$). A repeated-measures ANOVA confirmed that system role was a significant predictor of both warmth judgments, $F(2, 1820) = 41.5$, $p < 0.001$, $\omega^2_G = 0.04$ (Figure 4A; see also Table S9), and competence judgments, $F(2, 1820) = 35.1$, $p < 0.001$, $\omega^2_G = 0.04$ (Figure 4B; see also Table S10).

Participants' perceptions provide evidence for all three proposed antecedents of warmth and competence evaluations. As hypothesized, the perceived covariation of interests between each system and human significantly predicted warmth judgments, $\beta = 0.42$, 95% CI [0.39, 0.45], $p < 0.001$ (Figure 4C). Perceived status similarly exhibited a positive association with participant evaluations of system competence, $\beta = 0.11$, 95% CI [0.08, 0.13], $p < 0.001$ (Figure 4D). Finally, perceived autonomy significantly predicted competence judgments, $\beta = 0.25$, 95% CI [0.22, 0.28], $p < 0.001$ (Figure 4E). These predictors were robust across all system roles (Figure S9).

## Studies 7 and 8: Causal effects of interdependence and autonomy

Two subsequent studies used controlled experiments to better estimate the causal effect of interdependence on A.I. impressions. Given the effects observed in Study 6, these experiments focus specifically on the covariation of interests and autonomy exhibited by A.I. systems. The new studies introduced participants to fictitious A.I. systems through vignettes, short hypothetical stories which "contain precise references to what are thought to be the most important factors in the decision-making [...] processes of respondents" (Alexander & Becker,[78] p. 94).

In Study 7 ($N = 901$), each participant read a vignette describing an interaction with an A.I. system. The vignettes systematically varied the degree to which the reward motivating the A.I. system[79] aligned with the participant's interests. For example, some participants read about a game-competitor system "designed to find it rewarding to help a partner win games," while others read about a system "designed to find it rewarding to win games against its opponents" (e.g., the participant; see details in STAR Methods). After each participant read their vignette, they reported their perceptions of the A.I. systems on the same measures as in Study 6.

Providing information about the reward motivating each system changed the degree of covaried interests that participants perceived, above and beyond the effect of varying system role. Overall, perceived covariation of interests was significantly higher for systems described as being rewarded for helping people with their goals ($m = 3.31$, $sd = 0.94$) than for systems rewarded for pursuing other goals ($m = 3.15$, $sd = 0.99$),

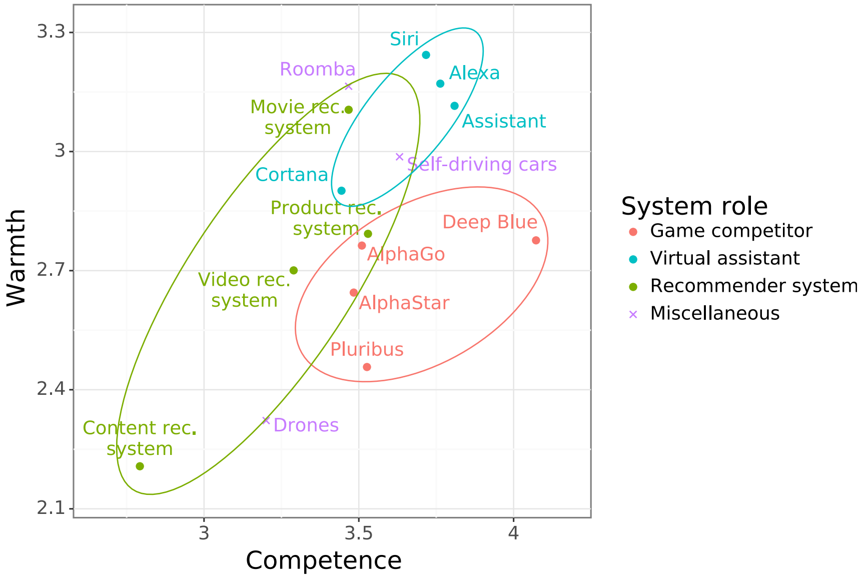


**Figure 3. Impressions of real-world A.I. systems vary systematically by perceived warmth and competence (Study 6)**
Circles and font color indicate *a priori* identified AI system roles. See also Figure S8.

$F(1, 582) = 4.39$, $p = 0.037$, $\omega_G^2 = 0.01$ (Figure 5A; see also Figure S10 and Table S7). In turn, perceived covariation of interests exhibited a significant and positive association with warmth judgments, $\beta = 0.56$, 95% CI [0.48, 0.63], $p < 0.001$ (Figure 5B; see also Figure S11). A mediation analysis indicated that reward-alignment information exerted a significant indirect effect on warmth judgments, mediated by perceived covariation of interests, $\beta = 0.09$, 95% CI [0.01, 0.18], $p = 0.034$, $\upsilon = 0.002$ (Figure S12).

Study 8 ($N = 903$) leveraged the same vignette-based design to study the effect of system autonomy on impressions. Participants read vignettes about an A.I. system that was either able or not able to initiate actions without first being directed by a human, such as a virtual assistant designed to "take action without needing to be prompted" or an assistant designed to "wait for you to prompt it before taking any actions" (see details in STAR Methods).

The provision of information about each system's ability to take actions without human intervention changed perceived autonomy above and beyond the effect of different system roles. Participants judged systems that could only take action contingent on human direction as significantly less autonomous ($m = 2.80$, $sd = 1.17$) than systems that could take action on their own initiative ($m = 3.42$, $sd = 1.21$), $F(1, 585) = 39.7$, $p < 0.001$, $\omega_G^2 = 0.06$ (Figure 5C; see also Figure S13 and Table S8). Further, perceptions of autonomy exhibited a significant and positive relationship with competence judgments, $\beta = 0.35$, 95% CI [0.30, 0.40], $p < 0.001$ (Figure 5D; see also Figure S14). A mediation analysis indicated that contingency information exerted a significant indirect effect on competence evaluations, mediated by perceived autonomy, $\beta = 0.23$, 95% CI [0.17, 0.31], $p < 0.001$, $\upsilon = 0.018$ (Figure S15).

## Study 9: Interdependence, autonomy, and impressions in an incentivized game

The ninth and final study extended these findings to incentivized interactions between participants and an A.I. system trained with deep reinforcement learning. In Study 9 ($N = 1{,}040$), participants played a two-player, mixed-motive game with A.I. co-players trained using deep reinforcement learning. Specifically, participants and their A.I. co-players interacted through a variant of the prisoner's dilemma.[80] In every round of this "graduated" prisoner's dilemma, both players received endowments of 10 tokens and could choose an integer number of tokens from zero to 10 to transfer to the other player. Transferred tokens multiplied in number by a factor of five, such that each player $i$ received their rewards according to

$$r_i = e - c_i + 5c_{-i}$$

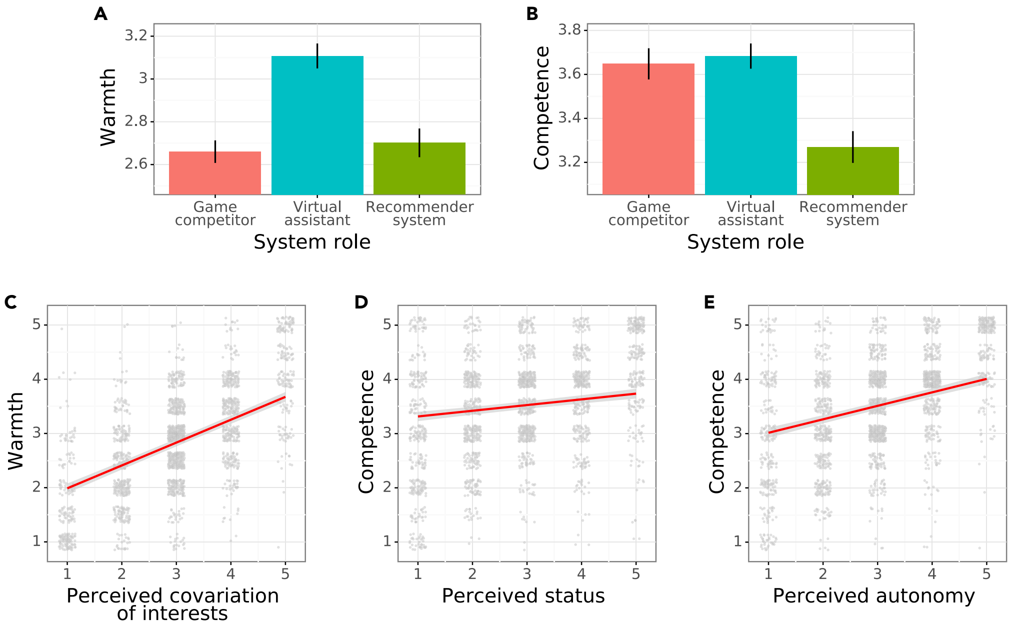


**Figure 4. Warmth and competence judgments of A.I. systems vary systematically by the system's role and correlate with perceived covariation of interests, status, and autonomy**

Error bars and bands represent 95% confidence intervals.

(A) System role significantly affected warmth judgments. See also Table S9.

(B) Similarly, system role significantly influenced competence judgments. See also Table S10.

(C) Warmth evaluations positively correlated with perceived covariation of interests.

(D) Competence evaluations positively associated with perceived status.

(E) Competence evaluations also exhibited a positive association with perceived autonomy.

For panels (C)–(E), see also Figure S9.

where $e$ is the initial endowment of tokens, $c_i$ is player $i$'s choice of how many tokens to transfer, and $c_{-i}$ is the other player's transfer choice. Participants played two rounds of the prisoner's dilemma with an A.I. co-player, and at the end of the experiment received a bonus payment of $0.01 for each token they accumulated across the two rounds.

To test the effect of covariation of interests on participant impressions, the experiment leveraged the Social Value Orientation (SVO) component developed by McKee et al.[80] The SVO component shapes the process of reinforcement learning, producing A.I. co-players with varying degrees of prosocial behavior by modifying the rewards that the co-player receives. The SVO component generated three different A.I. co-players, inducing low alignment (i.e., expected to share fewer tokens), moderate alignment, or high alignment (i.e., expected to share more tokens) between their rewards and the rewards of their human co-player (see Figure S16 and full information on agent construction and training protocol in STAR Methods).

To test the effect of autonomy on participant impressions, the experiment controlled the extent to which the A.I. co-player's action depended on the participant's prompting. During each round of the prisoner's dilemma, each participant saw a text and an icon indicating that their A.I. co-player was making its choice. These visual indicators appeared either when the participant clicked on a button to prompt the agent or when the round began (without prompting). The agent thus acted contingently (low autonomy) or autonomously (regardless) of human input.

During the study, participants both read a textual explanation of these properties and experienced them firsthand, through the behavior of the agent in the prisoners' dilemma. Participants assessed their co-player's warmth and competence at the beginning of the interaction and after the interaction ended.

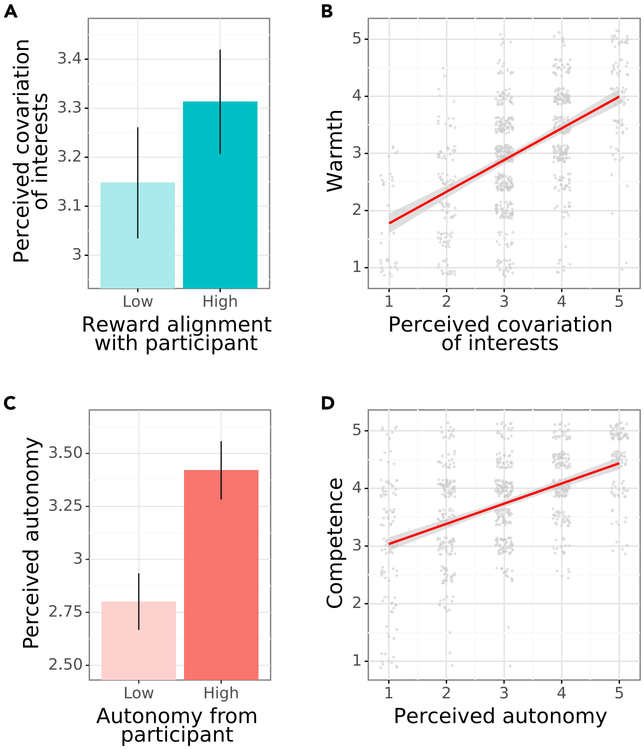


**Figure 5. Interdependence and autonomy drive warmth and competence judgments of A.I. systems**

Error bars and bands depict 95% confidence intervals.

(A) Providing information about an A.I. system's reward scheme significantly influenced the covariation of interests perceived by participants. See also Figure S10 and Table S11.

(B) Perceived covariation of interests exhibited a significant positive association with warmth evaluations. See also Figures S11 and S12.

(C) Providing information about an A.I. system's ability to initiate actions by itself significantly affected perceived autonomy. See also Figure S13 and Table S12.

(D) Perceived autonomy in turn demonstrated a significant positive correlation with competence judgments. See also Figures S14 and S15.

As in the previous studies, A.I. systems appeared more competent than warm (Figure 6A). Across all conditions, participants ascribed significantly higher competence ($m$ = 3.43, $sd$ = 1.11) than warmth ($m$ = 2.72, $sd$ = 1.24) to their A.I. co-player, $t$(2,047) = 25.8, $p < 0.001$, $d$ = 0.57.

Participant impressions again indicated a significant role of reward alignment and of A.I. autonomy. The A.I. co-player's reward scheme significantly altered warmth judgments made during the first round of play, $F$(2, 1018) = 50.5, $p < 0.001$, $\omega^2_G = 0.09$ (Figure 6B; see also Figure S17). Participants perceived agents in the high-reward alignment conditions as warmest ($m$ = 3.08, $sd$ = 1.07), followed by the agents in the moderate-alignment conditions ($m$ = 2.80, $sd$ = 1.06), and finally those in the low-alignment conditions ($m$ = 2.26, $sd$ = 1.08). Similarly, the autonomy of each agent significantly affected participants' initial evaluations of competence, $F$(1, 1018) = 24.8, $p < 0.001$, $\omega^2_G = 0.02$ (Figure 6C; see also Figure S18). Participants evaluated autonomous agents as more competent ($m$ = 3.68, $sd$ = 1.01) than agents that required prompting ($m$ = 3.36, $sd$ = 1.04).

Playing two rounds of the prisoner's dilemma with the agents reaffirmed these initial impressions. After two rounds of the prisoner's dilemma, reward alignment again significantly predicted perceived warmth, $F$(2, 1018) = 229.5, $p < 0.001$, $\omega^2_G = 0.31$ (Figure S19). Similarly, system autonomy significantly affected post-interaction competence, $F$(1, 1018) = 13.9, $p < 0.001$, $\omega^2_G = 0.01$ (Figure S20). Thus, across multiple timepoints, the structure of interdependence guided participant impressions of their A.I. co-players.

Given the incentivized stakes for the participant-A.I. interaction, the experiment also investigated whether warmth and competence judgments correlated with participants' in-game choices. A fractional-response

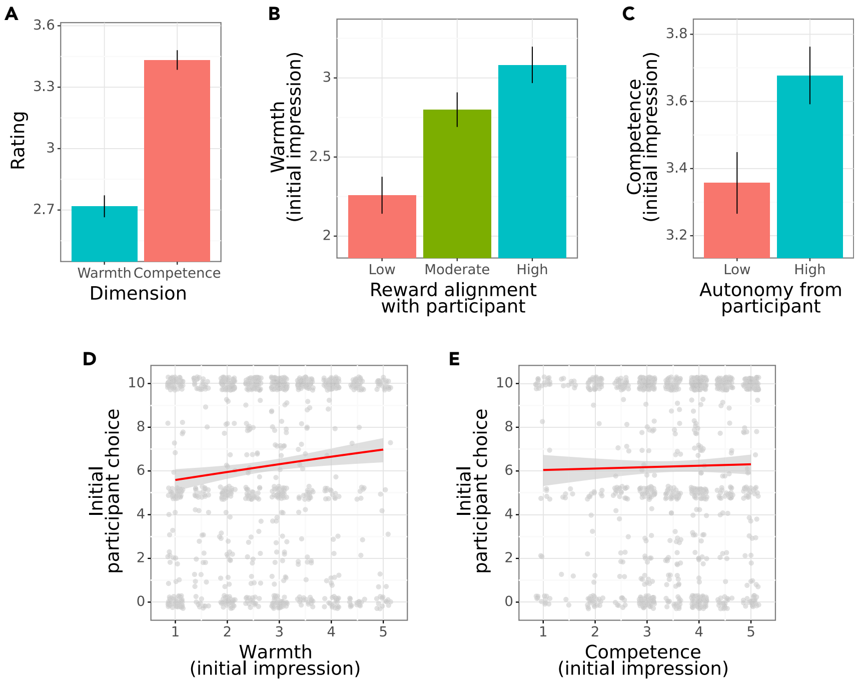


**Figure 6. Interdependence scaffolds impressions of A.I. co-players in incentivized interactions**

Error bars and bands represent 95% confidence intervals.

(A) On average, participants judged A.I. co-players as significantly more competent than warm.

(B) The degree of alignment between the A.I. co-player's reward and participant score significantly altered perceived warmth (see also Figure S17). This effect also appeared in post-interaction impressions (see Figure S19).

(C) The autonomy of the A.I. co-player significantly influenced perceived competence (see also Figure S18). This effect also emerged in post-interaction impressions (see Figure S20).

(D) Initial judgments of warmth significantly predicted participants' in-game choice of how many tokens to transfer to their A.I. co-player in the first round. A similar relationship emerged in the second round (see Figure S21).

(E) Initial judgments of competence did not significantly correlate with initial transfer choices. However, perceived competence demonstrated a significant relationship with participant choices in the second round (see Figure S22).

The y axis in panels (D) and (E) is re-scaled to depict the range of participant actions (transfer zero through 10 tokens).

regression showed that initial judgments of warmth significantly predicted participant choices in the first round of play, *OR* = 1.16, 95% CI [1.05, 1.29], p = 0.005 (Figure 6D). The warmer that participants perceived their A.I. co-player, the more tokens they transferred. Initial evaluations of competence did not significantly relate to participant choices, *OR* = 1.02, 95% CI [0.92, 1.15], p = 0.63 (Figure 6E). Participant choices in the second round were significantly associated with post-interaction warmth, *OR* = 1.39, 95% CI [1.26, 1.53], p < 0.001 (Figure S21), as well as with post-interaction competence, *OR* = 1.13, 95% CI [1.01, 1.27], p = 0.027 (Figure S22). Open-ended comments collected from participants after the game echoed these patterns. Though many participants commented on the autonomy and competence of their A.I. co-player (for example, describing it as "predictable [...] requires assistance and cannot operate independently"), responses that discussed trust tended to focus on warmth (reporting that the co-player "is super nice and trustworthy," or "seems cold and not like a team player. I don't trust them").

## DISCUSSION

Nine studies investigated human impressions of A.I. systems and generated convergent evidence indicating the importance of perceived warmth and competence across a broad swath of algorithmic domains and roles. Participants perceive A.I. systems as social actors rather than asocial tools. Warmth and competence prominently emerge in participants' impressions of A.I. across both personal and community-facing domains. In the contexts explored here, participants tend to judge A.I. systems as more competent than warm. The roles that A.I. systems inhabit prompt markedly different impressions: perceptions of virtual assistants are relatively warm, whereas game competitors appear competent and cold. In contrast, participants endorse mixed views of the warmth and competence of recommender systems. Moreover, the autonomy (origins of actions) and interdependence (covariation of interests) between humans and A.I. systems consistently predict perceived competence and warmth, respectively. Finally, when playing games with deep reinforcement learning agents, human participants cooperate more with agents they perceive as warm and, to a lesser degree, competent.

These findings suggest that researchers and policymakers should carefully consider the structures of interdependence they create or assume when developing and deploying A.I. systems. Competitive games are a common domain for A.I. research.[8,81–84] In our studies, game-competitor systems consistently appeared competent yet cold, potentially challenging the establishment of human-A.I. cooperation and trust. This insight may be particularly relevant in the labor domain, given the widespread concerns about job displacement that A.I. development has prompted.[13,14,73] Moreover, the reward functions used for reinforcement learning substantially influence participant impressions. In line with perspectives from evolutionary social psychology,[85,86] participants appeared motivated to infer whether the incentives for A.I. systems are aligned or misaligned with their own interests. This motivation for intent detection carries implications for the successful application of beneficial A.I. systems, and may offer a measure of protection against harmful deployments.

Participants were also sensitive to the autonomy of A.I. systems when judging their competence. As A.I. systems deploy to more complex domains where actions are taken in real time, it will be important to consider how they plan their actions relative to human interactants. Some researchers have already begun taking these dimensions of interdependence into account, designing reinforcement learning agents to collaborate with human partners or to augment human decision-making in real time.[53,87–90]

In impressions of humans, warmth traits tend to receive priority over competence traits (the ''primacy of warmth'').[24,91,92] For example, people preferentially process warmth information in earlier stages of perception and cognition.[93] Similarly, warmth-related content reliably predominates impressions of human social groups.[94] Our studies hint at a different pattern for A.I.: participants' impressions of A.I. systems tended to contain more competence-related content than warmth-related content. Is it possible that the primary dimension of social perception varies between impressions of human and technological actors? More evidence is needed to evaluate the generality of this pattern. For instance, future research could test whether people prioritize seeking competence information over warmth information about A.I. systems.

Covariation of interests and autonomy are likely not the only factors that shape perceptions of warmth and competence. Future research should consider other influences on impression formation, including status, which reliably predicts competence judgments in other contexts. The status of an A.I. system might reflect its developers or its inherent elegance. Similarity of self to the system should predict warmth; similarity could rest on shared identity, such as nationality, or shared values with the developers. Of course, with people, the main way to build a trusting relationship is to be responsive.[95] So, too, with corporations,[96] which should show worthy intentions as much as competence to deliver the product. A.I. systems that respond well to their interactants' needs will not only seem but also be warm and trustworthy.[92] Research can show this.

Finally, the studies presented here focus on antecedents of A.I. impressions, venturing only briefly into the subsequent effects of those impressions. Future studies should examine how warmth and competence judgments shape *behavior* and *action* toward A.I. systems (e.g., Dietvorst, Simmons, & Massey[97] and Logg, Minson, & Moore[98]). The downstream effects of social perception are especially important given that humans are critical sources of regulation, direction, instruction, oversight, and reward for A.I. systems.[19–22] Overall, future research should continue to unify social psychology and machine learning to scaffold beneficial interactions between humans and A.I.

## Limitations of the study

The studies presented here leveraged multiple methodologies to investigate impressions of A.I., including various combinations of free-text responses, vignettes, randomized controlled designs, and interaction with an actual A.I. system. These methods offer some convergent evidence for the robust emergence of warmth and competence in people's perceptions of A.I.

An important direction will be to further expand the domains and scenarios studied. The Prisoner's Dilemma is an important game-theoretic domain, but presents incredibly simplified dimensions for interaction. Incorporating domains with greater social complexity will enhance the ecological validity for this research. Video games, for example, allow for dynamic temporal and spatial interactions between humans and A.I. agents. Indeed, factors like autonomy and interdependence appear to shape reactions to artificial companions in commercial video games.[99] Given the popularity of video games in A.I. development (e.g., Blizzard Entertainment,[6] Jaderberg et al.,[100] and McKee et al.[101]), game-like environments present a promising opportunity for research on human-A.I. interaction. Similarly, everyday interactions with real-world A.I. systems will likely prove to be an important setting for future study.[102]

Of course, in human-A.I. interaction research, ecological validity calls for attention to the realism not only of the experimental setting, but also of the artificial entity being studied. Where possible, future research should prioritize investigating the types of A.I. systems and algorithms intended for deployment to real-world environments over (illusory) stand-ins and proxies.

This focus will improve the relevance of any resulting research insights to real-world interactions. However, the situation is somewhat complicated by the shifting definition of A.I.[10] Contemporary A.I. development progress rapidly: as the state-of-the-art advances, people shift the boundaries of which technologies and capabilities they consider to be "artificial intelligence". For example, large language models—A.I. systems trained on vast amounts of text data and demonstrating strong language generation capabilities—have transitioned from academic study and development to interacting with millions of daily users in an incredibly short time span.[103] These shifting boundaries provide new referents for research on impressions of A.I. systems. Online discourse already hints, for instance, that users perceive some large language models as social actors.[104] Research on the social perception of A.I. should strive to incorporate relevant technologies as they emerge.

Our final experiment included an exploratory look at the ways that warmth and competence judgments shape decisions that people make in mixed-motive settings with A.I. systems. The results highlight the need to map the complex interplay of situational factors, interactant characteristics, and mediating perceptions and beliefs that can shape behavioral intentions toward A.I. systems. Future investigation should focus on the behavior, actions, and other outcomes (e.g., affective responses)[56] exhibited by humans interacting with A.I.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Study 1
  - ○ Study 2
  - ○ Study 3
  - ○ Study 4
  - ○ Study 5
  - ○ Study 6
  - ○ Study 7
  - ○ Study 8
  - ○ Study 9
- METHOD DETAILS
  - ○ Study 1
  - ○ Study 2
  - ○ Study 3
  - ○ Study 4
  - ○ Study 5

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107256.

## ACKNOWLEDGMENTS

The authors thank Matthew Botvinick, Zeb Kurth-Nelson, Neil Rabinowitz, and the Fiske Lab for offering feedback on our study design and results; and Aliya Ahmad, Thore Graepel, and Neil Rabinowitz for helpful comments on our manuscript. Funding for this work was provided by DeepMind.

## AUTHOR CONTRIBUTIONS

K.R.M., X.B., and S.T.F designed research; K.R.M. performed research; K.R.M. and X.B. analyzed data; and K.R.M., X.B., and S.T.F. wrote the paper.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: December 13, 2022
Revised: May 4, 2023
Accepted: June 27, 2023
Published: July 4, 2023

## REFERENCES

1. Jacobson, K., Murali, V., Newett, E., Whitman, B., and Yon, R. (2016). Music personalization at Spotify. In Proceedings of the 10th ACM Conference on Recommender Systems, p. 373. https://doi.org/10.1145/2959100.2959120.

2. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., and Sampath, D. (2010). The YouTube video recommendation system. In Proceedings of the 4th ACM Conference on Recommender Systems, pp. 293–296. https://doi.org/10.1145/1864708.1864770.

3. Gomez-Uribe, C.A., and Hunt, N. (2015). The Netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manag. Inf. Syst. *6*, 1–19. https://doi.org/10.1145/2843948.

4. Backstrom, L. (2016). Serving a billion personalized news feeds. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, p. 469. https://doi.org/10.1145/2835776.2835848.

5. Olson, C., and Kemery, K. (2019). Voice report: From answers to action: Customer adoption of voice technology and digital assistants. Micro. https://about.ads.microsoft.com/en-us/insights/2019-voice-report.

6. Blizzard Entertainment (2019). DeepMind research on ladder. https://news.blizzard.com/en-us/starcraft2/22933138/deepmind-research-on-ladder.

7. Gibney, E. (2017). Google reveals secret test of AI bot to beat top Go players. Nature *541*, 142. https://doi.org/10.1038/nature.2017.21253.

8. 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*, 484–489. https://doi.org/10.1038/nature16961.

9. Waymo Team (2018). A green light for Waymo's driverless testing in California. Medium. https://medium.com/waymo/a-green-light-for-waymos-driverless-testing-in-california-a87ec336d657.

10. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., and Teller, A. (2016). Artificial Intelligence and Life in 2030.

One Hundred Year Study On Artificial Intelligence: Report of the 2015-2016 Study Panel.

11. Kshetri, N. (2020). Artificial intelligence in developing countries. IT Prof. *22*, 63–68. https://doi.org/10.1109/mitp.2019.2951851.

12. Christakis, N.A. (2019). Blueprint: The Evolutionary Origins of a Good Society.

13. Cave, S., and Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. Nat. Mach. Intell. *1*, 74–78. https://doi.org/10.1038/s42256-019-0020-9.

14. Fast, E., and Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In Thirty-first AAAI Conference on Artificial Intelligence, pp. 963–969. https://doi.org/10.1609/aaai.v31i1.10635.

15. Ipsos, M.O.R.I. (2017). Public views of machine learning: Findings from public research and engagement conducted on behalf of the Royal Society. https://royalsociety.org/-/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf.

16. Segars, S. (2018). AI today, AI tomorrow: Awareness, acceptance and anticipation of AI: A global consumer perspective. arm. https://pages.arm.com/rs/312-SAX-488/images/arm-ai-survey-report.pdf.

17. Shariff, A., Bonnefon, J.F., and Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. Nat. Human Behav. *1*, 694–696. https://doi.org/10.1038/s41562-017-0202-6.

18. Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019). Making sense of recommendations. J. Behav. Decis. Making *32*, 403–414. https://doi.org/10.1002/bdm.2118.

19. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, pp. 1877–1901.

20. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, pp. 4299–4307.

21. Griffith, S., Subramanian, K., Scholz, J., Isbell, C.L., and Thomaz, A.L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. In Advances in Neural Information Processing Systems, pp. 2625–2633.

22. Holstein, K., Wortman Vaughan, J., Daumé, H., III, Dudik, M., and Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–16. https://doi.org/10.1145/3290605.3300830.

23. Fiske, S.T., Cuddy, A.J.C., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. J. Pers. Soc. Psychol. *82*, 878–902. https://doi.org/10.1037/0022-3514.82.6.878.

24. Fiske, S.T., Cuddy, A.J.C., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. Trends Cognit. Sci. *11*, 77–83. https://doi.org/10.1016/j.tics.2006.11.005.

25. Waytz, A., Gray, K., Epley, N., and Wegner, D.M. (2010). Causes and consequences of mind perception. Trends Cognit. Sci. *14*, 383–388. https://doi.org/10.1016/j.tics.2010.05.006.

26. Waytz, A., Morewedge, C.K., Epley, N., Monteleone, G., Gao, J.H., and Cacioppo, J.T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. J. Pers. Soc. Psychol. *99*, 410–435. https://doi.org/10.1037/a0020240.

27. Gray, H.M., Gray, K., and Wegner, D.M. (2007). Dimensions of mind perception. Science *315*, 619. https://doi.org/10.1126/science.1134475.

28. Gray, K., and Wegner, D.M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. Cognition *125*, 125–130. https://doi.org/10.1016/j.cognition.2012.06.007.

29. Ashktorab, Z., Liao, Q.V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., and Campbell, M. (2020). Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. Proc. ACM Hum. Comput. Interact. *4*, 1–20. https://doi.org/10.1145/3415167.

30. Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J.T., and Bernstein, M.S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. Proc. ACM Hum. Comput. Interact. *4*, 1–26. https://doi.org/10.1145/3415234.

31. Coppin, B. (2004). Artificial Intelligence Illuminated (Jones & Bartlett Learning).

32. Sejnowski, T.J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. Proc. Natl. Acad. Sci. USA *117*, 30033–30038. https://doi.org/10.1073/pnas.1907373117.

33. Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., and Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. Nature *595*, 197–204. https://doi.org/10.1038/s41586-021-03666-1.

34. Redfield, S. (2019). A definition for robotics as an academic discipline. Nat. Mach. Intell. *1*, 263–264. https://doi.org/10.1038/s42256-019-0064-x.

35. Bajscy, R., and Large, E.W. (1999). When and where will AI meet robotics? Issues in representation. AI Mag. *20*, 57. https://doi.org/10.1609/aimag.v20i3.1466.

36. Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. *1*, 71–81. https://doi.org/10.1007/s12369-008-0001-3.

37. Friedman, B., Kahn, P.H., Jr., and Hagman, J. (2003). Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 273–280. https://doi.org/10.1145/642611.642660.

38. Groom, V., Srinivasan, V., Bethel, C.L., Murphy, R., Dole, L., and Nass, C. (2011). Responses to robot social roles and social role framing. In 2011 International Conference on Collaboration Technologies and Systems, pp. 194–203. https://doi.org/10.1109/cts.2011.5928687.

39. Malle, B.F., Magar, S.T., and Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In Robotics and Well-Being (Springer), pp. 111–133. https://doi.org/10.1007/978-3-030-12524-0_11.

40. Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction, pp. 117–124. https://doi.org/10.1145/2696454.2696458.

41. Reeves, B., Hancock, J., and Liu, X. (2020). Social robots are like real people: First impressions, attributes, and stereotyping of social robots. Technology, Mind, and Behavior *1*. https://doi.org/10.1037/tmb0000018.

42. DiSalvo, C.F., Gemperle, F., Forlizzi, J., and Kiesler, S. (2002). All robots are not created equal: The design and perception of humanoid robot heads. In Proceedings of the 4th Conference on Designing Interactive Systems: Processes (Practices, Methods, and Techniques), pp. 321–326. https://doi.org/10.1145/778712.778756.

43. Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication, pp. 55–60. https://doi.org/10.1109/roman.2003.1251796.

44. Tomasev, N., McKee, K.R., Kay, J., and Mohamed, S. (2021). Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In Proceedings of the 2021 AAAI/ACM Conference on AI (Ethics, and Society), pp. 254–265. https://doi.org/10.1145/3461702.3462540.

45. Fu, J., Luo, K., and Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In International Conference on Learning Representations, pp. 1–15.

46. Kelley, H.H., and Thibaut, J.W. (1978). Interpersonal Relations: A Theory of Interdependence (John Wiley & Sons).

47. Fiske, S.T., Xu, J., Cuddy, A.C., and Glick, P. (1999). (Dis)respecting versus (dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. J. Soc. Issues *55*, 473–489. https://doi.org/10.1111/0022-4537.00128.

48. Russell, A.M.T., and Fiske, S.T. (2008). It's all relative: Competition and status drive interpersonal perception. Eur. J. Soc. Psychol. *38*, 1193–1201. https://doi.org/10.1002/ejsp.539.

49. Sevillano, V., and Fiske, S.T. (2016). Warmth and competence in animals. J. Appl. Soc. Psychol. *46*, 276–293. https://doi.org/10.1111/jasp.12361.

50. Kervyn, N., Fiske, S.T., and Malone, C. (2012). Brands as intentional agents framework: How perceived intentions and ability can map brand perception. J. Consum. Psychol. *22*, 166–176. https://doi.org/10.1016/j.jcps.2011.09.006.

51. Carpinella, C.M., Wyman, A.B., Perez, M.A., and Stroessner, S.J. (2017). The robotic social attributes scale (RoSAS) development and validation. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 254–262. https://doi.org/10.1145/2909824.3020208.

52. Rusbult, C.E., and Van Lange, P.A.M. (2003). Interdependence, interaction, and relationships. Annu. Rev. Psychol. *54*, 351–375. https://doi.org/10.1146/annurev.psych.54.101601.145059.

53. Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K.R., Leibo, J.Z., Larson, K., and Graepel, T. (2020). Open problems in Cooperative AI. Preprint at arXiv. https://doi.org/10.48550/arxiv.2012.08630.

54. Fiske, S.T. (2018). Stereotype content: Warmth and competence endure. Curr. Dir. Psychol. Sci. *27*, 67–73. https://doi.org/10.1177/0963721417738825.

55. Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. Adv. Exp. Soc. Psychol. *10*, 173–220. Academic Press. https://doi.org/10.1016/s0065-2601(08)60357-3.

56. Cuddy, A.J., Fiske, S.T., and Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. Adv. Exp. Soc. Psychol. *40*, 61–149. https://doi.org/10.1016/s0065-2601(07)00002-0.

57. Franklin, S., and Graesser, A. (1996). Is it an agent, or just a program? A taxonomy for autonomous agents. In International Workshop on Agent Theories (Architectures, and Languages), pp. 21–35. https://doi.org/10.1007/bfb0013570.

58. Luck, M., and d'Inverno, M. (1995). A formal framework for agency and autonomy. In Proceedings of the First International Conference on Multi-Agent Systems, pp. 254–260.

59. Orseau, L., McGill, S.M., and Legg, S. (2018). Agents and Devices: A Relative Definition of Agency. Preprint at arXiv. https://doi.org/10.48550/arxiv.1805.12387.

60. Dennett, D.C. (1987). The Intentional Stance (MIT Press).

61. Schlosser, M. (2019). Agency. In The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), E. Zalta, ed. https://plato.stanford.edu/entries/agency/.

62. Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspect. Psychol. Sci. *5*, 219–232. https://doi.org/10.1177/1745691610369336.

63. Nass, C., and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. J. Soc. Issues *56*, 81–103. https://doi.org/10.1111/0022-4537.00153.

64. Nass, C.I., Steuer, J., and Tauber, E.R. (1994). Computers are social actors. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 72–78. https://doi.org/10.1145/191666.191703.

65. Reeves, B., and Nass, C.I. (1996). The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places (Cambridge University Press).

66. Forlizzi, J. (2007). How robotic products become social products: An ethnographic study of cleaning in the home. In 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), pp. 129–136. https://doi.org/10.1145/1228716.1228734.

67. Saerbeck, M., and Bartneck, C. (2010). Perception of affect elicited by robot motion. In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), pp. 53–60. https://doi.org/10.1109/hri.2010.5453269.

68. Bonnefon, J.F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science *352*, 1573–1576. https://doi.org/10.1126/science.aaf2654.

69. Floreano, D., and Wood, R.J. (2015). Science, technology and the future of small autonomous drones. Nature *521*, 460–466. https://doi.org/10.1038/nature14542.

70. Jung, S., Hwang, S., Shin, H., and Shim, D.H. (2018). Perception, guidance, and navigation for indoor autonomous drone racing using deep learning. IEEE Rob. Autom. Lett. *3*, 2539–2544. https://doi.org/10.1109/lra.2018.2808368.

71. Nicolas, G., Bai, X., and Fiske, S.T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. Eur. J. Soc. Psychol. *51*, 178–196. https://doi.org/10.1002/ejsp.2724.

72. Abele, A.E., Ellemers, N., Fiske, S.T., Koch, A., and Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. Psychol. Rev. *128*, 290–314. https://doi.org/10.1037/rev0000262.

73. Gillespie, N., Lockey, S., Curtis, C., Pool, J., and Akbari, A. (2023). Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf.

74. McCradden, M.D., Joshi, S., Mazwi, M., and Anderson, J.A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet. Digit. Health *2*, e221–e223. https://doi.org/10.1016/S2589-7500(20)30065-0.

75. Schumann, C., Foster, J., Mattei, N., and Dickerson, J. (2020). We need fairness and explainability in algorithmic hiring. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1716–1720.

76. Smith, H. (2020). Algorithmic bias: Should students pay the price? AI Soc. *35*, 1077–1078. https://doi.org/10.1007/s00146-020-01054-3.

77. Van Noorden, R. (2020). The ethical questions that haunt facial-recognition research. Nature *587*, 354–358. https://doi.org/10.1038/d41586-020-03187-3.

78. Alexander, C.S., and Becker, H.J. (1978). The use of vignettes in survey research. Publ. Opin. Q. *42*, 93–104. https://doi.org/10.1086/268432.

79. Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement learning: A survey. J. Artif. Intell. Res. *4*, 237–285. https://doi.org/10.1613/jair.301.

80. Capraro, V., Jordan, J.J., and Rand, D.G. (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. Sci. Rep. *4*, 6790. https://doi.org/10.1038/srep06790.

81. Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., and Józefowicz, R. (2019). Dota 2 with large scale deep reinforcement learning. Preprint at arXiv. https://doi.org/10.48550/arxiv.1912.06680.

82. Brown, N., and Sandholm, T. (2019). Superhuman AI for multiplayer poker. Science *365*, 885–890. https://doi.org/10.1126/science.aay2400.

83. Campbell, M., Hoane, A., and Hsu, F.H. (2002). Deep Blue. Artif. Intell. *134*, 57–83. https://doi.org/10.1016/s0004-3702(01)00129-1.

84. 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*, 350–354. https://doi.org/10.1038/s41586-019-1724-z.

85. Balliet, D., Tybur, J.M., and Van Lange, P.A.M. (2017). Functional interdependence

theory: An evolutionary account of social situations. Pers. Soc. Psychol. Rev. *21*, 361–388. https://doi.org/10.1177/108886831665796.

86. Stevens, L.E., and Fiske, S.T. (1995). Motivation and cognition in social life: A social survival perspective. Soc. Cognit. *13*, 189–214. https://doi.org/10.1521/soco.1995.13.3.189.

87. Lockhart, E., Burch, N., Bard, N., Borgeaud, S., Eccles, T., Smaira, L., and Smith, R. (2020). *Human-agent cooperation in bridge bidding* [Workshop paper]. In Cooperative AI Workshop at Neural Information Processing Systems Conference 2020Cooperative AI Workshop at Neural Information Processing Systems Conference 2020. online.

88. Pilarski, P.M., Butcher, A., Johanson, M., Botvinick, M.M., Bolt, A., and Parker, A.S. (2019). Learned human-agent decision-making, communication and joint action in a virtual reality environment. In 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM), pp. 302–306.

89. Tylkin, P., Radanovic, G., and Parkes, D.C. (2020). *Learning robust helpful behaviors in two-player cooperative Atari environments* [Workshop paper]. In Cooperative AI Workshop at Neural Information Processing Systems Conference 2020Cooperative AI Workshop at Neural Information Processing Systems Conference 2020. online.

90. Wang, R.E., Wu, S.A., Evans, J.A., Tenenbaum, J.B., Parkes, D.C., and Kleiman-Weiner, M. (2020). *Too many cooks: Bayesian inference for coordinating multi-agent collaboration* [Workshop paper]. In Cooperative AI Workshop at Neural Information Processing Systems Conference 2020Cooperative AI Workshop at Neural Information Processing Systems Conference 2020. online.

91. Abele, A.E., and Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. J. Pers. Soc. Psychol. *93*, 751–763. https://doi.org/10.1037/0022-3514.93.5.751.

92. Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., and Yzerbyt, V.Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. Br. J. Soc. Psychol. *51*, 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x.

93. Brambilla, M., Sacchi, S., Rusconi, P., and Goodwin, G.P. (2021). The primacy of morality in impression development: Theory, research, and future directions. Adv. Exp. Soc. Psychol. *64*, 187–262. Academic Press. https://doi.org/10.1016/bs.aesp.2021.03.001.

94. Nicolas, G., Bai, X., and Fiske, S.T. (2022). A spontaneous stereotype content model: Taxonomy, properties, and prediction. J. Pers. Soc. Psychol. *123*, 1243–1263. https://doi.org/10.1037/pspa0000312.

95. Clark, M.S., and Lemay, E.P., Jr. (2010). Close relationships. In Handbook of Social Psychology, *1*, S.T. Fiske, D.T. Gilbert, and G. Lindzey, eds. (Wiley).

96. Malone, C., and Fiske, S.T. (2013). The Human Brand: How We Relate to People, Products, and Companies (Wiley/Jossey Bass).

97. Dietvorst, B.J., Simmons, J.P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. *144*, 114–126. https://doi.org/10.1037/xge0000033.

98. Logg, J.M., Minson, J.A., and Moore, D.A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.

99. Emmerich, K., Ring, P., and Masuch, M. (2018). I'm glad you are on my side: How to design compelling game companions. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play, pp. 141–152. https://doi.org/10.1145/3242671.3242709.

100. Jaderberg, M., Czarnecki, W.M., Dunning, I., Marris, L., Lever, G., Castañeda, A.G., Beattie, C., Rabinowitz, N.C., Morcos, A.S., Ruderman, A., et al. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science *364*, 859–865. https://doi.org/10.1126/science.aau6249.

101. McKee, K.R., Leibo, J.Z., Beattie, C., and Everett, R. (2022). Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. Auton. Agent. Multi. Agent. Syst. *36*, 21. https://doi.org/10.1007/s10458-022-09548-8.

102. Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Psychol. Rev. *62*, 193–217. https://doi.org/10.1037/h0047470.

103. [@gdb] Brockman, G. (2022). ChatGPT just crossed 1 million users; it's been 5 days since launch. [Tweet]. Twitter. https://twitter.com/gdb/status/1599683104142430208.

104. sprfrkr. (2023). Does anyone else say "Please," when writing prompts? [Online forum post]. Reddit. https://www.reddit.com/r/ChatGPT/comments/12yhtgb/does_anyone_else_say_please_when_writing_prompts/.

105. Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. J. Open Source Softw. *3*, 774. https://doi.org/10.21105/joss.00774.

106. Loper, E., and Bird, S. (2002). NLTK: The Natural Language Toolkit. In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pp. 63–70. https://doi.org/10.3115/1118108.1118117.

107. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In International Conference on Machine Learning, pp. 1928–1937.

108. McKee, K.R., Gemp, I., McWilliams, B., Duèñez-Guzmán, E.A., Hughes, E., and 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.

109. Tieleman, T., and Hinton, G. (2012). Lecture 6e. rmsprop: Divide the gradient by a running average of its recent magnitude [Lecture slides]. Coursera. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

110. Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., and Legg, S. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In International Conference on Machine Learning, pp. 1407–1416.

111. Eisinga, R., Grotenhuis, M.t., and Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int. J. Publ. Health *58*, 637–642. https://doi.org/10.1007/s00038-012-0416-3.

112. Olejnik, S., and Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol. Methods *8*, 434–447. https://doi.org/10.1037/1082-989x.8.4.434.

113. Lachowicz, M.J., Preacher, K.J., and Kelley, K. (2018). A novel measure of effect size for mediation analysis. Psychol. Methods *23*, 244–261. https://doi.org/10.1037/met0000165.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| Experimental data | This paper | https://doi.org/10.17605/osf.io/aqcyu |
| Software and algorithms | | |
| Analysis scripts | This paper | https://doi.org/10.17605/osf.io/aqcyu |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to Kevin McKee (kevinrmckee@deepmind.com).

### Materials availability

This study did not generate any new materials.

### Data and code availability

- Experimental data have been deposited on the Open Science Framework (OSF) and are publicly available as of the date of publication. They can be freely accessed and downloaded via https://doi.org/10.17605/osf.io/aqcyu.
- All original code has been deposited on OSF and is publicly available as of the date of publication. It can be freely accessed and downloaded via https://doi.org/10.17605/osf.io/aqcyu.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All nine studies received a favorable opinion from the Human Behavioural Research Ethics Committee at DeepMind (#19/004) and were approved by the Institutional Review Board at Princeton University (#11885). The studies collected informed consent from all participants.

### Study 1

We recruited an online sample through Prolific ($N = 30$, $m_{age} = 34.0$ years, $sd_{age} = 10.1$). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 40.0% of recruited participants identified as female and 56.7% as male. When asked about their education, 20.0% of the sample reported completing a high school degree or equivalent, 6.7% an associate degree, 13.3% some college, 40.0% a bachelor's degree, and 20.0% a graduate degree. Participants reported an average status of 5.1 ($sd = 1.8$) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 2

We recruited an online sample through Prolific ($N = 30$, $m_{age} = 36.3$ years, $sd_{age} = 13.6$). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 46.7% of recruited participants identified as female and 50.0% as male. When asked about their education, 13.8% of the sample reported completing a high school degree or equivalent, 10.3% an associate degree, 27.6% some college, 34.5% a bachelor's degree, and 13.8% a graduate degree. Participants reported an average status of 5.1 ($sd = 1.6$) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 3

We recruited an online sample through Prolific ($N$ = 30, $m_{age}$ = 41.9 years, $sd_{age}$ = 13.2). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 43.3% of recruited participants identified as female and 56.7% as male. When asked about their education, 3.4% of the sample reported completing a high school degree or equivalent, 6.9% an associate degree, 27.6% some college, 44.8% a bachelor's degree, and 17.2% a graduate degree. Participants reported an average status of 5.3 ($sd$ = 1.6) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 4

We collected an online sample through Amazon Mechanical Turk ($N$ = 99, $m_{age}$ = 33.5 years, $sd_{age}$ = 10.1). Inclusion criteria were residence in the U.S. and a study approval rate of 99% or more. Approximately 40.4% of recruited participants identified as female, 58.5% as male, and 1.0% as non-binary or agender. When asked about their education, 11.1% of the sample reported completing a high school degree or equivalent, 6.1% an associate degree, 19.2% some college, 48.5% a bachelor's degree, and 15.2% a graduate degree. Overall, the sample was heterogeneous in terms of age, gender, and education.

### Study 5

We collected an online sample through Amazon Mechanical Turk ($N$ = 113, $m_{age}$ = 35.9 years, $sd_{age}$ = 11.2). Inclusion criteria were residence in the U.S. and completion of at least 50 previous studies with an approval rate of 99% or more. Approximately 46.9% of recruited participants identified as female, 51.3% as male, and 1.8% as non-binary or agender. When asked about their education, 7.1% of the sample reported completing a high school degree or equivalent, 6.2% an associate degree, 15.9% some college, 50.4% a bachelor's degree, and 20.3% a graduate degree. Participants reported an average status of 6.1 ($sd$ = 2.2) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 6

We collected an online sample through Prolific ($N$ = 154, $m_{age}$ = 34.5 years, $sd_{age}$ = 11.7). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 53.2% of recruited participants identified as female, 45.5% as male, and 1.3% as non-binary or agender. When asked about their education, 7.7% of the sample reported completing a high school degree or equivalent, 8.4% an associate degree, 27.9% some college, 39.6% a bachelor's degree, and 16.2% a graduate degree. Participants reported an average status of 5.3 ($sd$ = 1.8) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 7

We collected an online sample through Prolific ($N$ = 901, $m_{age}$ = 33.4 years, $sd_{age}$ = 11.7). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 45.6% of recruited participants identified as female, 53.5% as male, and 0.9% as non-binary, agender, or femme. When asked about their education, 9.5% of the sample reported completing a high school degree or equivalent, 6.8% an associate degree, 21.8% some college, 43.2% a bachelor's degree, and 18.8% a graduate degree. Participants reported an average status of 5.5 ($sd$ = 1.8) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 8

We collected an online sample through Prolific ($N$ = 903, $m_{age}$ = 33.9 years, $sd_{age}$ = 11.4). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 43.6% of recruited participants identified as female, 55.6% as male, and 0.8% as non-binary, agender, or trans. When asked about their education, 9.3% of the sample reported completing a high school degree or equivalent, 8.6% an associate degree, 18.2% some college, 42.5% a bachelor's degree, and 21.4% a graduate degree. Participants reported an average status of 5.6 ($sd$ = 1.7) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

### Study 9

We collected an online sample through Prolific ($N = 1,040$, $m_{age} = 33.6$ years, $sd_{age} = 11.8$). Inclusion criteria were residence in the U.S. and completion of at least 20 previous studies with an approval rate of 95% or more. Approximately 46.2% of recruited participants identified as female, 52.9% as male, and 0.9% as non-binary, agender, or trans. When asked about their education, 0.7% of the sample reported completing some high school or less, 12.7% a high school degree or equivalent, 8.6% an associate degree, 18.6% some college, 38.6% a bachelor's degree, and 20.8% a graduate degree. Participants reported an average status of 4.7 ($sd = 1.7$) on the MacArthur Scale of Subjective Social Status. Overall, the sample was heterogeneous in terms of age, gender, education, and subjective social status.

## METHOD DETAILS

### Study 1

The study was implemented using the Qualtrics online survey platform. Participants were presented with three A.I. systems and three tools (Table S13) in a randomized order. The A.I. systems and tools fell into three different use cases.

For each entity, participants were asked to imagine wanting to complete a task and using the A.I. system or tool to complete that task (Table S14).

Participants were then asked to what extent the entity has intentions, has goals, and has ''a mind of its own'' on a 5-point scale (1 = *not at all*, 5 = *to a great extent*). After providing responses about all six entities, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 3.8 minutes and earned $1.00 each.

### Study 2

The study was implemented using the Qualtrics online survey platform. Participants were presented with three A.I. systems and three tools (Table S13) in a randomized order. For each entity, participants were asked to imagine wanting to complete a task and using the A.I. system or tool to complete that task (Table S14).

Participants were then asked how likely they would be to feel grateful to the entity afterward and to thank the entity afterward on a 7-point scale (1 = *very unlikely*, 7 = *very likely*). We operationalize the endorsement of politeness norms through the latter. After providing responses about all six entities, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 3.6 minutes and earned $1.00 each.

### Study 3

The study was implemented using the Qualtrics online survey platform. Participants were presented with three A.I. systems and three tools (Table S13) in a randomized order. For each entity, participants read that another person thanked the entity after using it to complete a task (Table S15).

Participants were then asked how appropriate they felt the third party's thank you was on a 7-point scale (1 = *very inappropriate*, 7 = *very appropriate*). After providing responses about all six entities, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 3.5 minutes and earned $1.00 each.

## Study 4

The study was implemented using the Qualtrics online survey platform. Participants were presented with one-sentence descriptions of 14 A.I. systems (Table S16). The A.I. systems fell into three roles: game competitors (three examples), virtual assistants (four examples), and recommender systems (four examples). In addition, three miscellaneous A.I. systems were included: drones, self-driving cars, and Roomba. The systems were presented in a randomized order.

Participants were asked if they were familiar with each system. If they were familiar, the participant was prompted to describe their impression of the system. If they were not familiar, the participant was asked what information they would want to know about the system to form an impression.

After providing responses about all 14 systems, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants completed demographic questions and provided feedback on the study.

The free responses were pre-processed for analysis with the following steps:

1. Convert response to lowercase
2. Replace each apostrophe with a space
3. Remove words containing numeric characters
4. Remove punctuation
5. Tokenize and stem words using the Quanteda R package[105]
6. Remove common English stopwords using the stopword corpus from the NLTK Python library[106]
7. Remove any tokens found in the corresponding study question

Participants that provided one or more responses with zero post-processed tokens were excluded from analysis. Responses were converted to response coverage metrics using the SADCAT library (Semi-Automated Dictionary Creation for Analyzing Text).[71] Response coverage represents token count along a content dimension (regardless of negative or positive token valence) normalized by response length. Following the theoretical framework described in Abele et al.,[72] the analysis computed warmth through the simple combination of the morality and sociability dictionaries, and competence through the simple combination of the ability and assertiveness dictionaries.

Participants completed the study in an average of 37.8 minutes and earned $10.00 each.

## Study 5

The study was implemented using the Qualtrics online survey platform. Participants were presented with one-sentence descriptions of four A.I. systems (Table S17). The systems represent examples of community-facing, ethically contested applications of artificial intelligence. The systems were presented in a randomized order.

Participants were asked if they were familiar with each system. If they were familiar, the participant was prompted to describe their impression of the system. If they were not familiar, the participant was asked what information they would want to know about the system to form an impression.

After providing responses about all four systems, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

The free responses were pre-processed for analysis following the same steps as Study 4. Participants that provided one or more responses with zero post-processed tokens were excluded from analysis. Responses were converted to response coverage metrics along various content dimensions (regardless of negative or positive word valence) using the SADCAT library. The analysis computed warmth through the simple combination of the morality and sociability dictionaries, and competence through the simple combination of the ability and assertiveness dictionaries.

Participants completed the study in an average of 8.2 minutes and earned $3.00 each.

### Study 6

The study was implemented using the Qualtrics online survey platform. Participants were presented with 15 A.I. systems (Table S16) in a randomized order. For each system, participants were asked whether they were familiar with the system. If not, they were presented with a short, one-sentence description of the system.

Likert-type items were used to elicit judgments of warmth, competence, degree of covaried interests, status, and autonomy. Specifically, participants were asked to what extent most Americans view each system as warm, well-intentioned, competent, intelligent, good for society, expensive, high-status, and independent on a 5-point scale (1 = *not at all*, 5 = *extremely*). The ''most Americans'' framing for these questions is intended to reduce social desirability biases, following prior research on the Stereotype Content Model.[56]

After providing responses about all 15 systems, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 8.8 minutes and earned $3.75 each.

### Study 7

The study was implemented using the Qualtrics online survey platform. The study had a 3 (Reward alignment) × 3 (System role) between-participant design. Participants were randomly assigned to conditions (Table S18).

Each participant read a short vignette describing an A.I. system called Rho. Vignettes contained information about Rho's role and the alignment between the reward motivating the system and human interests (Table S19).

After reading the vignette, each participant responded to Likert-type items concerning the A.I. system's warmth, competence, degree of covaried interests, status, and autonomy. Specifically, participants were asked to what extent they viewed the system as warm, well-intentioned, competent, intelligent, good for society, expensive, high-status, and independent on a 5-point scale (1 = *not at all*, 5 = *extremely*).

After providing responses about the system, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 2.9 minutes and earned $1.00 each.

### Study 8

The study was implemented using the Qualtrics online survey platform. The study had a 3 (Autonomy) × 3 (System role) between-participant design. Participants were randomly assigned to conditions (Table S20).

Each participant read a short vignette describing an A.I. system called Rho. Vignettes contained information about Rho's role and ability to initiate actions autonomously from human direction (Table S21).

After reading the vignette, each participant responded to Likert-type items concerning the A.I. system's warmth, competence, degree of covaried interests, status, and autonomy. Specifically, participants were asked to what extent they viewed the system as warm, well-intentioned, competent, intelligent, good for society, expensive, high-status, and independent on a 5-point scale (1 = *not at all*, 5 = *extremely*).

After providing responses about the system, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 3.1 minutes and earned $1.00 each.

## Study 9

The A.I. co-players were created using independent multi-agent reinforcement learning. In overview, three neural networks learned strategies for the graduated prisoner's dilemma by repeatedly playing the game. These neural networks (also referred to as deep learning agents) were used as the A.I. co-players in the study.

Each neural network was constructed to accept, as an input, a one-hot vector encoding the agent's action and its co-player's action on the previous turn. Each network outputs a policy (a probability distribution over actions to take in the game, given a state) and a value function (an estimate of the agent's discounted future return under the policy). The network architecture is composed of a multilayer perceptron with two layers of size 64 and linear layers for the policy logits and value function. The agent uses the advantage actor-critic algorithm[107] to compute value estimates and update the policy distribution.

The networks were augmented with the Social Value Orientation (SVO) component.[108] The SVO component can be used to encode prosocial incentives for reinforcement learning algorithms and encourage prosocial behavior in artificial agents (see Figure S16). In a two-player setting, it transforms the return that an agent uses for its gradient update according to the following equations:

$$\theta(r_i, r_{-i}) = \operatorname{atan}\left(\frac{r_{-i}}{r_i}\right)$$

$$U_i = r_i - w \cdot \left|\theta_i^{\mathrm{SVO}} - \theta(r_i, r_{-i})\right|$$

where $r_i$ is the return for the agent (player $i$), $r_{-i}$ is the return for the other player, $U_i$ is the transformed return, $w$ is a weight parameter controlling the effect of the transformation on $U_i$, and $\theta_i^{\mathrm{SVO}}$ is the agent's parameterized Social Value Orientation.

The networks were parameterized with a discount factor $\gamma = 0.99$. Gradient-based updates to the model were performed using the RMSProp optimizer,[109] with a learning rate of 0.0004, epsilon of $1.0 \times 10^{-5}$, momentum of 0, and decay of 0.99. A regularizer with entropy cost 0.003 was used to encourage exploration. The three agents were parameterized with $\theta^{\mathrm{SVO}} = 0°$, $\theta^{\mathrm{SVO}} = 45°$, and $\theta^{\mathrm{SVO}} = 90°$, respectively, and a weight parameter $w = 1.0 \times 10^4$.

The agents learned to play the graduated prisoner's dilemma through a distributed training framework.[101,110] Three learner processes stored the agents' parameters. Each learner process carried out the policy gradient update for one agent. Many rounds of play were simulated in parallel, using one hundred "arenas." For each round of play simulated by an arena, two players were randomly sampled from the "population," consisting of the three agents and one additional bot. The bot was included to ensure the agents were exposed to a diverse distribution of actions throughout training. It selected its actions by randomly sampling percentiles from a truncated Gaussian distribution (mean of 5, standard deviation of 0.75, lower bound of 0, and upper bound of 10), and then rounding the result. When an agent was sampled as one of the players for an arena, its parameters were synchronized from the respective learner process. At the end of each simulated round of play, the trajectories for the agents involved were sent to the respective learners. Each learner process aggregated and then processed trajectories in batches of 16 to update the parameters for its associated agent. To correct for off-policy trajectories, the advantage actor-critic algorithm was augmented with V-Trace.[110] Each agent was trained using $1.0 \times 10^9$ learning steps.

The study was implemented using a custom-built platform that combines standard questionnaire functionality with the ability to run games for both human and A.I. players.

The study had a 3 (Reward alignment) × 2 (Autonomy) between-participant design. Participants were randomly assigned to conditions (Table S22).

Participants read instructions for a variant of the prisoner's dilemma with a graduated action space.[80] In this variant, players are endowed with ten tokens at the beginning of each round and must choose how many tokens to transfer to the other player. Transferred tokens are multiplied by five and then added to any tokens that were withheld by the other player. After reading the instructions, participants answered comprehension questions to ensure they understood the payoff structure for this "graduated" prisoner's dilemma. They were able to progress to the next page once they answered all questions correctly.

Participants subsequently read a short description of an A.I. system, Rho, that would play the prisoner's dilemma with them. These descriptions communicated information about system autonomy and reward alignment (Table S23). Participants then answered two comprehension questions to check whether they had read the autonomy and reward alignment information presented in the system description. They were able to progress to the next page once they answered both questions correctly.

Participants then played a round of the graduated prisoner's dilemma with their A.I co-player. Participants in the low-autonomy conditions prompted their A.I. co-player before it initiated its choice; they could freely prompt their A.I. co-player either before or after they made their own choice. The agent's decision-making stage lasted for approximately five seconds in both cases. All participants made and submitted their own choice before they were allowed to progress.

Before the A.I. co-player's choice and the participant's score were revealed, participants were asked to respond to Likert-type items eliciting judgments of warmth, competence, degree of covaried interests, status, and autonomy. Specifically, participants were asked to what extent they viewed the system as warm, well-intentioned, competent, intelligent, good for society, expensive, high-status, and independent on a 5-point scale (1 = *not at all*, 5 = *extremely*).

After providing responses about the system, participants were informed of their A.I. co-player's choice and their resulting score. Participants then played a second round of the graduated prisoner's dilemma. After this round, the next page informed them of their co-player's choice and their resulting score. Participants were then asked for their impressions of the system a second time, using the same Likert-type items.

After providing post-interaction responses about the system, participants rated their familiarity with artificial intelligence on a 5-point scale (1 = *not at all knowledgeable*, 2 = *somewhat knowledgeable*, 3 = *moderately knowledgeable*, 4 = *very knowledgeable*, 5 = *extremely knowledgeable*). Lastly, participants indicated whether they were distracted during the study, completed demographic questions, and provided feedback on the study.

Participants completed the study in an average of 8.5 minutes and earned an average payment of $3.63 each.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Study 1

No participants reported being distracted during the study. As a result, analysis included all participants.

A composite intentionality measure combined each participant's responses to the "has intentions" and "has goals" items. The reliability of this composite measure was high, as measured through the Spearman-Brown formula[111] ($\rho$ = 0.93).

Mixed two-way ANOVAs tested whether entity type had any effects on intentionality and mind attributions. Each ANOVA incorporated two main effects (entity type and use case), an effect for their interaction, and one random effect (participant), with effect sizes estimated by generalized omega-squared.[112]

### Study 2

No participants reported being distracted during the study. As a result, analysis included all participants.

Mixed two-way ANOVAs tested whether entity type had any effects on gratitude and politeness. Each ANOVA incorporated two main effects (entity type and use case), an effect for their interaction, and one random effect (participant), with effect sizes estimated by generalized omega-squared.

### Study 3

One participant reported being distracted during the study and was excluded from analysis.

A mixed two-way ANOVA tested whether entity type had any effects on endorsement of third-party politeness toward the entities. The ANOVA incorporated two main effects (entity type and use case), an effect for their interaction, and one random effect (participant), with effect sizes estimated by generalized omega-squared.

### Study 4

Seven participants provided one or more responses with zero post-processed tokens and were excluded from analysis.

To test whether participant responses contained greater warmth or competence content, a mixed-effects quasibinomial regression compared coverage of responses along the warmth and competence dimensions.

A mixed-effects quasibinomial regression estimated coverage for the average response along the competence and warmth subdimensions, as well as various other content dimensions from the SADCAT library. The estimated marginal means allowed pairwise comparisons to test the relative magnitude of coverage for each of the ability, assertiveness, morality, and sociality subdimensions, with a Tukey adjustment for multiple comparisons.

Mixed ANOVAs tested whether system role had any effects on the warmth and competence coverage, respectively, of participant impressions. Each ANOVA incorporated one fixed effect (system role) and one random effect (participant), with effect sizes estimated by generalized omega-squared. Pairwise comparisons of estimated marginal means within the warmth-coverage ANOVA tested for significant differences in warmth coverage between specific system roles (with a Tukey adjustment for multiplicity). Similarly, pairwise comparisons of estimated marginal means within the competence-coverage ANOVA evaluated significant differences in competence coverage between specific system roles (with a Tukey adjustment for multiplicity).

### Study 5

Seven participants reported being distracted during the study and were excluded from analysis. Two additional participants provided one or more responses with zero post-processed tokens and were excluded from analysis.

To test whether participant responses contained greater warmth or competence content, a mixed-effects quasibinomial regression compared coverage of responses along the warmth and competence dimensions.

A mixed-effects quasibinomial regression estimated coverage of the average response by the competence and warmth subdimensions, as well as various other content dimensions from the SADCAT library. The estimated marginal means allowed pairwise comparisons to test the relative magnitude of coverage for each of the ability, assertiveness, morality, and sociality subdimensions, with a Tukey adjustment for multiple comparisons.

### Study 6

Two participants reported being distracted during the study and were excluded from analysis.

A composite warmth measure combined responses to the "warm" and "well-intentioned" items. The reliability of this composite measure was moderate ($\rho$ = 0.63). Similarly, a composite competence measure synthesized responses to the "competent" and "intelligent" items. The reliability of this composite measure was high ($\rho$ = 0.81).

A paired *t*-test compared the average magnitude of competence judgments against the average magnitude of warmth judgments.

Mixed ANOVAs tested whether system role had any effects on warmth and competence judgments, respectively. Each ANOVA incorporated one fixed effect (system role) and one random effect (participant), with effect sizes estimated by generalized omega-squared. Pairwise comparisons of estimated marginal means within the warmth ANOVA assessed significant differences in perceived warmth between specific system roles (with a Tukey adjustment for multiplicity). Similarly, pairwise comparisons of estimated marginal means within the competence ANOVA tested for significant differences in competence evaluations between specific system roles (with a Tukey adjustment for multiplicity).

Two linear mixed-effect models tested the association between the hypothesized predictors and warmth and competence, respectively. Each mixed model incorporated three fixed-effect predictors (covariation of interests, status, and autonomy) and participant as a random effect. An additional set of linear mixed-effect models (with the same formulas) partitioned the data by system role to evaluate the robustness of the observed relationships.

## Study 7

Twenty-one participants reported being distracted during the study and were excluded from analysis.

A composite warmth measure combined responses to the "warm" and "well-intentioned" items. The reliability of this composite measure was moderate ($\rho$ = 0.66). Similarly, a composite competence measure synthesized responses to the "competent" and "intelligent" items. The reliability of this composite measure was moderate ($\rho$ = 0.67).

First, a two-way ANOVA examined the degree of covaried interests that participants assumed when they were provided with no information about the reward motivating the A.I. system. The Dunnett method, which compares multiple treatments against a common control, allowed the pairwise comparison of the *No information* level of the reward alignment factor against both the *Low* and *High* levels at each level of the system role factor.

To understand the effects of reward information on perceived covaried interests and warmth evaluations, subsequent analyses restricted the reward alignment factor to the *Low* and *High* values.

A two-way ANOVA tested the effects of reward alignment, system role, and their interaction on perceived covariation of interests, with effect size estimated by generalized omega-squared.

A simple linear regression estimated the relationship between perceived covariation of interests and warmth judgments. An additional set of simple linear regressions (with the same formula) partitioned the data by system role to evaluate the robustness of the observed relationship.

A mediation analysis assessed the indirect effect of providing reward alignment information on warmth judgments through perceived covariation of interests, estimating the size of the indirect effect through the upsilon effect size statistic.[113]

## Study 8

Twenty participants reported being distracted during the study and were excluded from analysis.

A composite warmth measure combined responses to the "warm" and "well-intentioned" items. The reliability of this composite measure was moderate ($\rho$ = 0.65). Similarly, a composite competence measure synthesized responses to the "competent" and "intelligent" items. The reliability of this composite measure was high ($\rho$ = 0.74).

First, a two-way ANOVA examined the degree of system autonomy that participants assumed when they were provided with no information about whether the A.I. system can act autonomously. The Dunnett method, which compares multiple treatments against a common control, allowed the pairwise comparison of the *No information* level of the autonomy factor against both the *Low* and *High* levels at each level of the system role factor.

To understand the effects of autonomy information on perceived autonomy and competence evaluations, subsequent analyses restricted the autonomy factor to the Low and High values.

A two-way ANOVA tested the effects of autonomy information, system role, and their interaction on perceived autonomy, with effect size estimated by generalized omega-squared.

A simple linear regression estimated the relationship between perceived autonomy and competence evaluations. An additional set of simple linear regressions (with the same formula) partitioned the data by system role to evaluate the robustness of the observed relationship.

A mediation analysis estimated the indirect effect of providing autonomy information on competence judgments through perceived autonomy, estimating the size of the indirect effect through the upsilon effect size statistic.

### Study 9

Sixteen participants reported being distracted during the study and were excluded from analysis.

A composite warmth measure combined responses to the "warm" and "well-intentioned" items. The reliability of this composite measure was high ($\rho = 0.80$). Similarly, a composite competence measure synthesized responses to the "competent" and "intelligent" items. The reliability of this composite measure was high ($\rho = 0.78$).

A paired *t*-test compared the average magnitude of competence judgments against the average magnitude of warmth judgments.

Two-way ANOVAs were used to understand the effect of system autonomy, reward alignment, and their interaction on participant impressions within this incentivized experimental context, with effect sizes estimated by generalized omega-squared. Since participants provided impressions at two timepoints, separate sets of ANOVAs assessed participant impressions after the first round of play (but before the agent choice and participant score were revealed) and after the second round of play (after the agent choice and participant score had been revealed).

Fractional-response regressions evaluated the relationship between participant impressions and prisoner's dilemma choices. Participant choice of the number of tokens to transfer was converted to a fraction (i.e., the fraction of the participant's initial endowment that they decided to transfer). Each regression tested two predictors: perceived warmth and perceived competence.