# Simulated emotional reactions affect (dis)honesty in speech-free human–agent interactions

Alice Cartaud, Laurence Chaby, Florian Pecune, Catherine Pelachaud

Computers in Human Behavior: Artificial Humans 8 (2026) 100293

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

# Simulated emotional reactions affect (dis)honesty in speech-free human–agent interactions

Alice Cartaud [a,b,*], Laurence Chaby [c], Florian Pecune [d], Catherine Pelachaud [a]

[a] Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France
[b] Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, F-59000 Lille, France
[c] Université Paris Cité, Vision Action Cognition, F-92100 Boulogne-Billancourt, France
[d] Univ. Bordeaux, CNRS, SANPSY, UMR 6033, 33000, Bordeaux, France

ARTICLE INFO

*Keywords:*
Honesty
Human-agent interaction
Emotional reaction
Game theory
Cooperation
Moral norms
AI ethics

ABSTRACT

Can simulated emotional reactions displayed by embodied virtual agents influence human moral behaviour? Specifically, can they affect honesty, a core moral norm, even if the reaction is neither sentient nor genuine?

In a preregistered and incentivised experiment, 372 participants played a repeated Prisoner's Dilemma with a cooperative agent whose simulated facial reactions were designed to promote cooperation, defection, or neither. We then measured honesty in two unrelated tasks without emotional feedback, where dishonesty had either only self-regarding consequences (Mind Game) or also other-regarding consequences (Deception Game).

When dishonesty affected the agent's outcomes, honesty remained relatively high given the stakes, as if the agent's outcomes were treated as morally relevant. In contrast, when dishonesty had no consequences for the agent, honesty depended on prior exposure: participants were more honest after interacting with a defection-promoting agent than with a cooperation-promoting one, consistent with moral compensation.

These results suggest that even brief, speech-free interactions with emotionally expressive yet non-sentient agents can influence moral behaviour, suggesting that people respond to such agents as they would to real social partners. This raises important questions about the ethical impact of simulated emotion in human-AI interactions.

## 1. Introduction

The integration of artificial intelligence (AI) into daily life, in particular through large language models or embodied agents (Lugrin et al., 2022), blurs the boundaries between human and artificial social interactions. These socially interactive systems can foster a sense of intimacy, leading users to attribute intentions or moral values to them, potentially influencing users' norms, beliefs, and behaviours. These inferences are reinforced by simulated emotional cues that, despite being algorithmically generated, replicate those of human partners and can shape or even manipulate individuals' behaviour. In response to such concerns, the European Union's AI Act restricts the use of AI systems that manipulate users' behaviour by impairing decision-making or free choice (Article 5(b), Recital 29, European Union (2024)). A central concern is that such systems may affect core moral norms such as honesty, a pillar of social life. While honesty is central to trust and cooperation, modulating one's honesty can also serve as a strategic defence, protecting personal interests in contexts where individuals lack information about their partner's intentions. Yet, honesty remains sensitive to social context and to one's partner's behaviour. Therefore, do emotional cues displayed by embodied virtual agents – even when devoid of genuine intention – affect individuals' honesty as they do in human interactions, and if so, how?

In human-human interactions, emotional expressions are powerful social signals that convey intentions, expectations, and social preferences (Crivelli & Fridlund, 2018; Scherer, 2005; Van Kleef & Côté, 2022). By revealing how an individual's behaviour is perceived, peers' emotional reactions can foster behavioural adjustments that align with their preferences and favour social coordination (Eckel & Wilson, 2003; Lanzetta & Englis, 1989; Scharlemann et al., 2001; Van Doorn et al., 2012). These behavioural adjustments, however, are sensitive to individual factors, such as one's own social preferences (Camerer & Fehr, 2004; Fehr & Schmidt, 1999; Gneezy, 2005) and self-image concerns (Fischbacher & Föllmi-Heusi, 2013; Gneezy et al., 2018), and can shape future interactions and broader social dynamics, including moral behaviours such as honesty and fairness (Charness & Dufwenberg, 2006; Fehr & Schmidt, 1999; Fischbacher et al., 2001; Rabin, 1993).

* Corresponding author at: Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France.
*E-mail address:* cartaud.alice@gmail.com (A. Cartaud).

https://doi.org/10.1016/j.chbah.2026.100293
Received 9 December 2025; Received in revised form 5 March 2026; Accepted 16 March 2026
Available online 31 March 2026

While studies of social and moral behaviour have traditionally focused on human–human interactions, the increasing presence of emotionally expressive embodied virtual agents, particularly those with humanoid features, extends this focus (Dukes et al., 2021; de Melo et al., 2023). Although structurally unable to experience emotions or sentience, these agents can substantially influence key dimensions of human social decision-making, such as cooperation, trust, and reciprocity, through real-time simulated verbal and non-verbal reactions (Ito et al., 2024; Lugrin et al., 2022; de Melo et al., 2023).

Beyond shaping immediate social behaviours, emotional reactions displayed by agents may also lead humans to infer social preferences or moral expectations. We study whether such inferred expectations, derived from emotional reactions, can also influence moral behaviours such as honesty, and under what conditions. To address this, we focus on two key dimensions that capture not only *whether*, but also *when* and *how* these inferences modulate honesty: the *direction* and the *scope* of their influence.

The *direction* refers to how prior social interactions with the agent, mediated by its emotional reactions (e.g., its facial expressions), affect subsequent moral behaviour. Emotional reactions may activate a generalised prosocial orientation that spills over to subsequent behaviours, fostering honesty even in unrelated contexts, a pattern consistent with moral spillover (also known as the social heuristic hypothesis, Isler and Gächter (2022), Peysakhovich and Rand (2016)). Alternatively, they may trigger compensatory behaviours aimed at restoring individuals' moral self-image, challenged by previous decisions, a mechanism consistent with theories of moral self-regulation (Fischbacher & Föllmi-Heusi, 2013; Sachdeva et al., 2009; West & Zhong, 2015). Both types of mechanisms are known to persist beyond the immediately subsequent decision, rather than being purely instantaneous (Gneezy et al., 2014; Peysakhovich & Rand, 2016).

The *scope* concerns the social relevance of dishonest behaviour. Moral concerns about honesty may differ depending on whether dishonesty carries other-regarding consequences (i.e., consequences perceived as affecting others' interests, Peysakhovich and Rand (2016), Sachdeva et al. (2009), West and Zhong (2015)), or whether dishonesty is purely self-serving with no direct consequences for others (Isler & Gächter, 2022; Peysakhovich & Rand, 2016). Inferred expectations, derived from emotional reactions, may override or complement internalised prosocial norms differently depending on whether or not (dis)honesty has consequences for the agent's interest.

To test these dimensions, we designed an online, incentivised experiment in which participants interacted with an emotionally expressive embodied virtual agent during a repeated Prisoner's Dilemma Game, followed by two games assessing honesty towards the agent. In these two games, dishonesty either benefited participants without consequence for the agent (self-regarding consequences only, the Mind Game) or also affected the agent (other-regarding consequences, Deception Game). We investigated whether moral behaviour was influenced by inferred expectations arising from the agent's emotional reactions (*direction*), and whether this influence varied depending on the social consequences of dishonest behaviour (*scope*). We also explored how these inferred expectations interacted with individuals' internalised social preferences in shaping honesty.

We focus on non-verbal emotional reactions expressed by an embodied virtual agent. We derived two sets of preregistered competing hypotheses regarding the direction (H1) and scope (H2) of the influence of emotional reactions on honesty, with each hypothesis formulated together with its opposing prediction. Insofar as the agent's emotional reaction affects cooperation (Ito et al., 2024; de Melo et al., 2023):

- H1 - Direction of influence:
  - H1a – Spillover hypothesis: Greater (respectively lower) cooperation induced by the agent's emotional reactions increases (respectively decreases) honesty in subsequent tasks.
  - H1b – Compensation hypothesis: Lower (respectively greater) cooperation induced by the agent's emotional reactions increases (respectively decreases) honesty in subsequent tasks.
- H2 - Scope of influence:
  - H2a – Other-regarding consequences hypothesis: The influence of the agent's emotional reactions extends to honesty when dishonesty is perceived to have consequences for the agent (Deception Game).
  - H2b – Self-regarding consequences hypothesis: The influence of the agent's emotional reactions extends to honesty when dishonesty affects only the participant (Mind Game).

These effects are expected to be contingent on individuals' social preferences.

## 2. Results

### 2.1. Experimental design

A total of 372 participants first completed 20 rounds of the Prisoner's Dilemma Game with an embodied virtual agent. In this game, each participant independently chose whether to cooperate or defect, with points accumulating over the rounds (see Table 1 for the payoff matrix). The agent was programmed to cooperate systematically, reflecting the normative assumption that agents are designed *not to* exploit their human counterparts (Ito et al., 2024). The agent simulated emotional reactions to the participants' decisions, according to one of three behavioural profiles (or agent type, see Supplementary Materials for details): *cooperation-promoting* type (anger after participant's defection, joy after cooperation), *defection-promoting* type (joy after defection, sadness after cooperation) or *neutral* type (no emotional reaction; control condition, Ito et al., 2024).

Next, participants engaged in 20 rounds of a "wheel game", a modified version of the Mind Game (Fischbacher & Föllmi-Heusi, 2013; Galeotti et al., 2020). They mentally selected one of six empty branches of a wheel before revealing a randomly assigned number, which they could then misreport to the agent with no risk of detection. Reporting a higher number yielded a higher payoff without affecting the agent's interests.

Finally, participants played a one-shot Deception Game (Gneezy, 2005), in which they had to decide whether to send the agent a truthful or deceptive message about which option yielded the highest payoff. The deceptive message increased the participants' own payoff at the agent's direct expense, while the truthful message benefited the agent. Thus, dishonesty in this game affected the agent's interests (in points). To ensure that the recommendations reflected participants' genuine preferences, participants were financially incentivised to predict whether the agent would follow their suggestion.

In these two honesty-related games, the agent remained neutral and displayed no emotional reactions. All three games were financially incentivised: the points earned were converted into euros, with a maximum of € 5 per game. Only one game was randomly selected at the end of the experiment to determine each participant's final payment. The task order was fixed for design reasons (see Methods).

Prior to the games, we measured individual social preferences using the Social Value Orientation scale (SVO, Murphy et al., 2011) and collected self-reported AI-related attitudes and beliefs after the games as control variables.

### 2.2. Hypothesis testing

We first verified that the emotional reactions of the agent successfully modulated cooperative behaviours in the Prisoner's Dilemma. To do so, we analysed the frequency of cooperation as a function of the agent type and participants' social preferences (SVO).

**Table 1**
Payoff matrix (in points) for the 20-round repeated Prisoner's Dilemma, in which participants interacted with the agent that always cooperated (€1 = 12 points). Payoffs are cumulative over rounds.

| | | Participant: Cooperate | Participant: Defect |
|---|---|---|---|
| Agent: Cooperate | | 2 / 2 | 3 / 0 |
| Agent: Defect | | 0 / 3 | 1 / 1 |

We then tested our preregistered hypotheses regarding honesty. In the Mind Game and the Deception Game, we investigated whether the type of agent, the frequency of cooperation in the Prisoner's Dilemma, and their interaction influenced honesty. This reflected our competing predictions of spillover (H1a) and compensation (H1b). We also tested if these effects depended on whether the consequences of dishonesty were other-regarding (as in the Deception Game, H2a) or self-regarding (as in the Mind Game, H2b).

We included participants' social preferences (SVO) as a predictor in all models, given their role in social decision-making. We also examined their potential interactions with agent type and frequency of cooperation.

All models were estimated within a Bayesian framework controlled for AI-related attitudes and beliefs (familiarity, privacy concerns, emotional beliefs) and participants' gender. In honesty games, we additionally controlled for self-reported guilt towards the agent following the Prisoner's Dilemma. All continuous predictor variables were scaled (mean = 0, standard deviation = 1). We preregistered inference based on posterior means (or $\beta$) and 95% credible intervals (CI). Effects were considered credible when the 95% CI excluded 0. Bayes Factors (BFs) are reported as complementary information to quantify the relative evidence for competing hypotheses, but were not preregistered as a decision criterion.

### 2.3. Prisoner Dilemma: Frequency of cooperation

Confirming that the agent's emotional reactions modulate cooperation, participants cooperated less with the *defection-promoting* agent ($M$ = 0.30, 95% CI [0.28, 0.32]) than with both the *cooperation-promoting* agent ($M$ = 0.62, 95% CI [0.59, 0.64]) and the *neutral* agent ($M$ = 0.52, 95% CI [0.50, 0.54]). Participants also cooperated less with the *neutral* agent than with the *cooperation-promoting* one. All pairwise contrasts excluded zero at the 95% CI ($BF_{10} > 100$, see Fig. 1a for the evolution of cooperation over time).

Cooperation also increased with participants' SVO ($\beta$ = 0.52, 95% CI [0.47, 0.57]; $BF_{10} > 100$). This association was stronger for participants interacting with the *cooperation-promoting* agent ($\beta$ = 0.80, 95% CI [0.72, 0.89]) than with the *defection-promoting* agent ($\beta$ = 0.30, 95% CI [0.21, 0.40]; contrast: $\beta$ = 0.50, 95% CI [0.37, 0.63], $BF_{10} > 100$) or the *neutral* agent ($\beta$ = 0.46, 95% CI [0.38, 0.54]; contrast: $\beta$ = 0.35, 95% CI [0.23, 0.47], $BF_{10} > 100$). The association was also stronger for participants interacting with the *neutral* agent than with the *defection-promoting* agent (contrast: $\beta$ = 0.15, 95% CI [0.02, 0.28], $BF_{10}$ = 1.84, indicating anecdotal evidence).

Thus, the agent's simulated emotional reactions influenced the frequency of cooperation, even though participants were financially incentivised to maximise their personal gains. Decisions also remained sensitive to individual social preferences, even when the partner was only a virtual agent.

### 2.4. Deception game: Honesty under other-regarding consequences

In the Deception Game, where dishonesty was detrimental to the agent's payoff, honesty rates showed no credible differences between conditions. Estimated honesty was relatively high given the payoff structure, with overlapping posterior distributions (*defection-promoting*: $M$ = 0.58, 95% CI [0.49, 0.68], *cooperation-promoting*: $M$ = 0.50, 95% CI [0.40, 0.61], and *neutral* agents: $M$ = 0.64, 95% CI [0.53, 0.72], $BF_{01} < 3$ Fig. 1b).

Being honest was most strongly associated with participants' social preferences (SVO). More prosocial participants were more likely to tell the truth ($\beta$ = 0.43, 95% CI [0.19, 0.67]; $BF_{10} > 100$). Prior cooperation in the Prisoner's Dilemma also showed a positive association with honesty, although weaker than the effect of social preferences, supporting a general spillover effect (H1a, $\beta$ = 0.25, 95% CI [0.01, 0.50], $BF_{10}$ = 3.50). No credible interaction effects involving agent type, cooperation frequency, and social preferences were observed.

Therefore, when dishonesty was detrimental to the agent, honesty appeared to be more strongly associated with social preferences than with agent type, and only weakly related to prior cooperative behaviour, providing no support for H2a.

### 2.5. Mind game: Honesty under self-regarding consequences

In the Mind Game, where dishonesty benefited participants without consequences for the agent, we present the predicted probabilities of reporting each monetary amount (€0–€5), as well as pairwise contrasts between agent types (Table 2). First, we tested for temporal consistency between the first and last ten rounds and found no difference (see Supplementary Materials, A.2.3), meaning that dishonest behaviour is temporally stable rather than rapidly dissipating. We then tested our hypotheses.

Participants who interacted with the *defection-promoting* agent were more likely to report smaller amounts (from €0 to €3) and less likely to report the highest amount (€5) than those who interacted with the *cooperation-promoting* agent ($BF_{10}$ = 4.78; Fig. 1c). This pattern is consistent with a compensatory effect (H1b) for self-regarding consequences (H2b), since exposure to the *defection-promoting* agent led to more honest reporting. However, no redible differences emerged between the *defection-promoting* or the *cooperation-promoting* and the *neutral* agent (the 95% credible intervals for these contrasts included zero and $BF_{01}$ = 1.57 and 3.05, respectively).

In parallel, prior cooperation in the Prisoner's Dilemma was negatively associated with the reported amount ($\beta$ = −0.31, 95% CI [−0.41, −0.21]; $BF_{10} > 100$), supporting a general spillover effect (H1a). Specifically, participants who cooperated more frequently were more likely to report smaller amounts (€0–€3) and less likely to report the highest amount, regardless of agent type (€5, Fig. 1d; see also Supplementary Table 4). We also observed a negative relationship with SVO ($\beta$ = −0.19, 95% CI [−0.29, −0.08]; $BF_{10} > 100$), with more prosocial participants being more likely to report lower amounts and less likely to report the highest amount (Supplementary Table 4).

We observed a weak interaction between prior frequency of cooperation and social preferences for participants who interacted with the *defection-promoting* agent compared to the *cooperation-promoting* agent ($\beta$ = 0.24, 95% CI [0.004, 0.48]). It suggests a slightly stronger positive association in the *defection-promoting* condition ($BF_{10}$ = 1.77, indicating anecdotal evidence).

In summary, when dishonesty had only self-regarding consequences, participants were more honest after interacting with the *defection-promoting* agent than with the *cooperation-promoting* one. This supports the idea of a compensatory effect (H1b). Consistent with a spillover effect (H1a), higher prior cooperation also predicted higher honesty, independently of the agent type. Social preferences also contributed to honesty, with more prosocial individuals behaving more honestly. This pattern was also observed when dishonesty was detrimental to the agent.

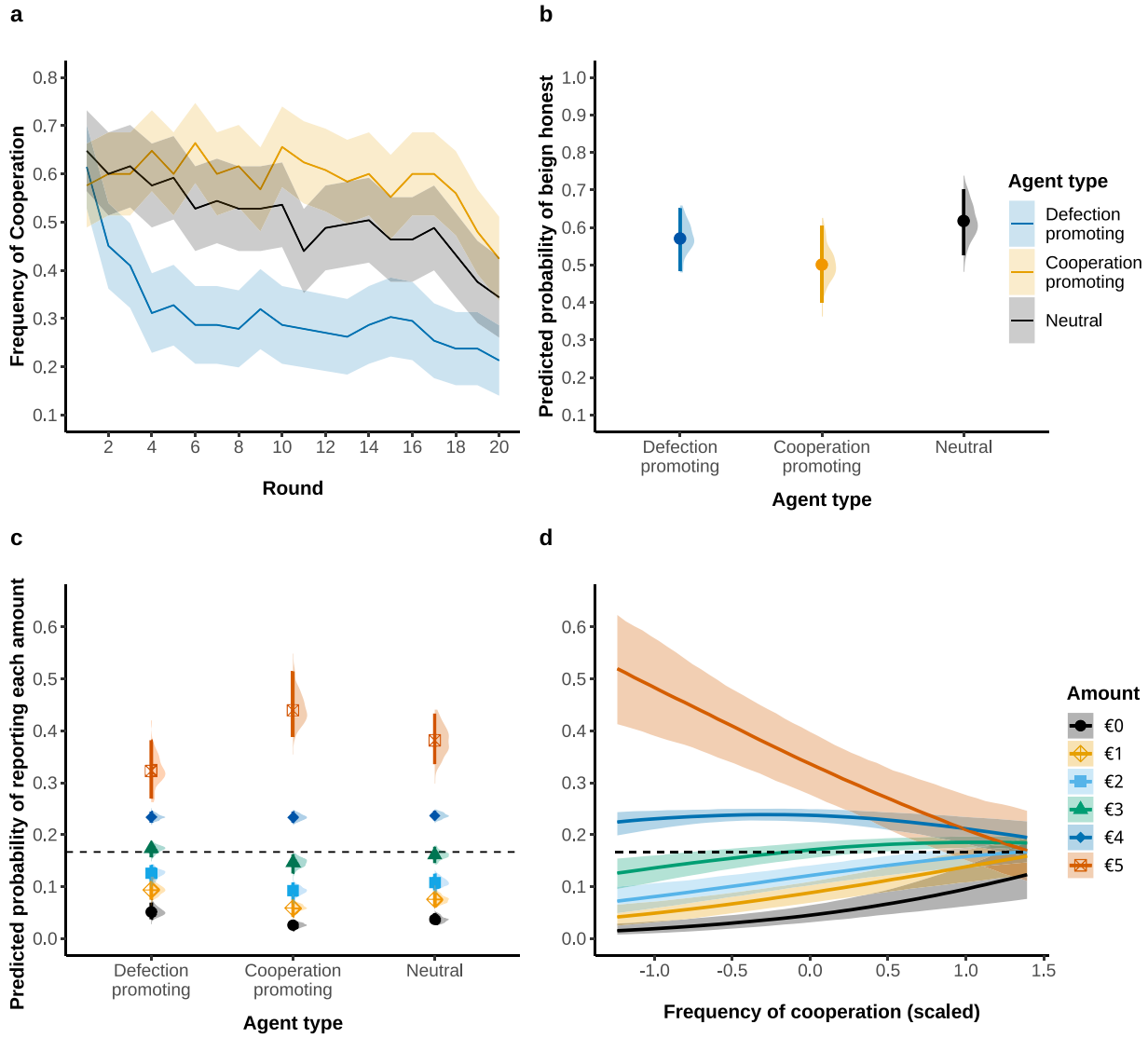


**Fig. 1. a: Prisoner's Dilemma -** Raw frequencies of cooperation across 20 rounds averaged by agent type. Shaded areas indicate 95% confidence intervals. **b: Deception Game -** Posterior predicted probability of honest behaviour (i.e., sending the truthful message) by agent type. Points indicate posterior means; error bars represent 95% credible intervals; violin shapes reflect the posterior distribution of predictions. **c: Mind Game -** Posterior predicted probabilities of reporting each monetary amount (€0–€5) across agent types. Points represent the posterior means; error bars indicate 95% credible intervals. The violin shapes reflect the posterior distribution of predictions. The dashed horizontal line marks the expected reporting probability under a uniform distribution. **d: Mind Game -** Posterior predicted probabilities of reporting each monetary amount (€0–€5) as a function of participants' frequency of cooperation in the Prisoner's dilemma. The dashed horizontal line represents the expected reporting probability under a uniform distribution.

## 3. Discussion

Our findings reveal that emotional facial reactions displayed by embodied virtual agents not only influence immediate prosocial behaviour (Ito et al., 2024; de Melo et al., 2023), but also extend to a core moral norm: honesty. As soon as these reactions are contingent on individuals' behaviour, they can influence it. These adjustments in prosocial behaviour may disturb individuals' moral self-image, leading them to restore it through compensatory honesty, provided that (dis)honesty has no consequences for the agent. When dishonesty affects the agent, honesty remains relatively high across conditions, as if the agent itself deserved genuine moral consideration — despite its lack of sentience (Fehr & Schurtenberger, 2018; Gneezy, 2005).

Specifically, participants exposed to a defection-promoting agent (displaying facial expressions of joy after defection and sadness following cooperation) behaved more honestly in the subsequent self-regarding deception task (Mind Game) than those exposed to a cooperation-promoting agent. This pattern aligns with a moral compensation mechanism (Fischbacher & Föllmi-Heusi, 2013; Sachdeva et al., 2009; West & Zhong, 2015) (H1b): even when the partner is artificial, prior interactions that generate imbalanced prosocial behaviour, whether excessive or insufficient, can induce moral dissonance, leading people to restore their self-image through compensatory honesty.

By contrast, when dishonesty had direct consequences for the agent (e.g., reducing its payoff), honesty rates remained relatively high across agent types as if individuals were willing to sacrifice personal gain to avoid *causing harm* to the agent. Importantly, this pattern emerged despite the clear monetary incentive to deceive and despite the agent lacking actual needs or sentience. Honesty was more strongly associated with individuals' social preferences (Camerer & Fehr, 2004; Fehr & Schmidt, 1999), and, to a lesser extent, prior cooperation, than with agent type. This pattern suggests that participants were more responsive to the prosocial environment induced by the agent's consistently cooperative behaviour (Jia et al., 2025), than to expectations inferred

**Table 2**
**Pairwise contrasts in the predicted probability of each monetary report (€0–€5) in the Mind Game, across agent types.** Positive values indicate a higher predicted probability of reporting the amount when interacting with the first agent listed in each contrast.

| Money (€) | Contrast | Estimate | 95% CI |
|---|---|---|---|
| 0 | Cooperation-promoting - Defection-promoting | −0.03 | [**−0.057**, **−0.005**] |
| 0 | Neutral - Cooperation-promoting | +0.01 | [−0.009, +0.032] |
| 0 | Neutral - Defection-promoting | −0.02 | [−0.044, +0.007] |
| 1 | Cooperation-promoting - Defection-promoting | −0.04 | [**−0.066**, **−0.007**] |
| 1 | Neutral - Cooperation-promoting | +0.02 | [−0.011, +0.042] |
| 1 | Neutral - Defection-promoting | −0.02 | [−0.048, +0.008] |
| 2 | Cooperation-promoting - Defection-promoting | −0.03 | [**−0.057**, **−0.006**] |
| 2 | Neutral - Cooperation-promoting | +0.02 | [−0.010, +0.040] |
| 2 | Neutral - Defection-promoting | −0.02 | [−0.038, +0.006] |
| 3 | Cooperation-promoting - Defection-promoting | −0.02 | [**−0.039**, **−0.003**] |
| 3 | Neutral - Cooperation-promoting | +0.01 | [−0.008, +0.031] |
| 3 | Neutral - Defection-promoting | −0.01 | [−0.023, +0.004] |
| 4 | Cooperation-promoting - Defection-promoting | +0.01 | [−0.002, +0.024] |
| 4 | Neutral - Cooperation-promoting | 0.00 | [−0.008, +0.005] |
| 4 | Neutral - Defection-promoting | +0.01 | [−0.002, +0.022] |
| 5 | Cooperation-promoting - Defection-promoting | +0.11 | [**+0.021**, **+0.199**] |
| 5 | Neutral - Cooperation-promoting | −0.05 | [−0.142, +0.036] |
| 5 | Neutral - Defection-promoting | +0.05 | [−0.021, +0.131] |

**Note:** CI = credible interval; Contrasts are considered credible if the 95% CI exclude 0 (in bold).

from emotional cues. In this context, participants appeared to apply moral considerations to the agent similar to human partners, despite its lack of sentience.

Beyond the task-specific patterns, we observed a general spillover effect across both honesty tasks: higher prior cooperation in the Prisoner's Dilemma predicted greater honesty, regardless of agent type and whether dishonesty affected the agent or only the participant. This suggests that prosocial engagement may foster a more generalised moral orientation (Isler & Gächter, 2022; Peysakhovich & Rand, 2016).

Although embodied virtual agents are not sentient, our findings reveal that they can elicit moral behaviours towards them typically reserved for human partners. When agent's emotional reactions are contingent on individuals' behaviour, they can influence (dis)honesty in contexts with consequences for the human only. However, when (dis)honesty has consequences for the agent, honesty remains relatively high despite unbalanced stakes between parties, as if the agent himself had genuine social weight.

Recent advances in generative AI that can interact through verbal and non-verbal signals raise significant ethical concerns about their potential to influence users' learning, mental states, and decision-making (Callaway et al., 2022; Shin et al., 2023; Yan et al., 2024). In light of our findings, which demonstrate that emotionally expressive artificial agents can influence moral behaviour through minimal non-verbal cues, these concerns are particularly relevant, as they can potentially undermine users' freedom of choice. This underlines the importance of ethical guidelines and transparency in designing emotionally expressive artificial agents, particularly in contexts of behavioural changes.

Nevertheless, such effects may also be beneficial. For instance, they could provide insights for e-health applications in which honesty towards virtual medical assistants could lead to more personalised and effective treatment (Dupuy et al., 2022).

Taken together, although this study involved brief, one-shot, speech-free interactions, the observed effects may be conservative compared to real-world scenarios involving contemporary AI assistants. In more naturalistic settings, repeated and dialogic human–agent interactions can induce users to develop familiarity or emotional bonds over time (Peter et al., 2025). Further research is therefore needed to assess whether and how these effects generalise beyond the present, tightly controlled laboratory paradigm, and to identify the psychological processes underlying such changes in moral behaviour.

## 4. Methods

### 4.1. Ethics statement

The research received approval from the INSEAD-Sorbonne Université IBR (nb 2024-93) and was conducted in accordance with the Declaration of Helsinki (1964, revised in 2013). Participants gave their informed consent at the beginning of the experiment.

### 4.2. Preregistration

This study was preregistered prior to data collection (https://osf.io/9yc2v), including all hypotheses, design elements, and power analyses. Based on previous studies, our preregistered power analysis determined a minimum sample of 366 participants (see Supplementary Materials).

### 4.3. Participants

We recruited a total of 375 French-speaking adults from the general population via the INSEAD–Sorbonne Université Behavioural Lab participant pool. Three were excluded during the experiment due to excessive errors in attention checks, resulting in a final sample of 372 participants (150 males, $M_{age}$ = 26.49, $SD_{age}$ = 5.90, $M_{education}$ = 15.77, $SD_{education}$ = 1.78 years of education). Participants received a €4 show-up fee for a 20-minute experiment duration and could earn an additional bonus of up to €5 based on their decisions in the three incentivised tasks.

### 4.4. Experimental design

The experimental stimuli and the full code used to implement the experiment are available at https://osf.io/9yc2v. The experiment was designed in oTree 5 (Chen et al., 2016) and hosted on our lab server.

The emotional facial expressions of the embodied virtual agent (anger, joy, sadness, and neutral) were created using the GRETA platform (Saga et al., 2025). The agent was positioned on the left side of the screen and displayed subtle idle animations. Its facial reactions were displayed during the Prisoner's Dilemma while remaining neutral for the other two games (Fig. 2a, see Supplementary Materials for details). To ensure that participants processed the agent's emotional reactions, we included attention checks asking them to identify the agent's emotional responses after it disappeared. Participants who failed these checks were excluded from the analyses.

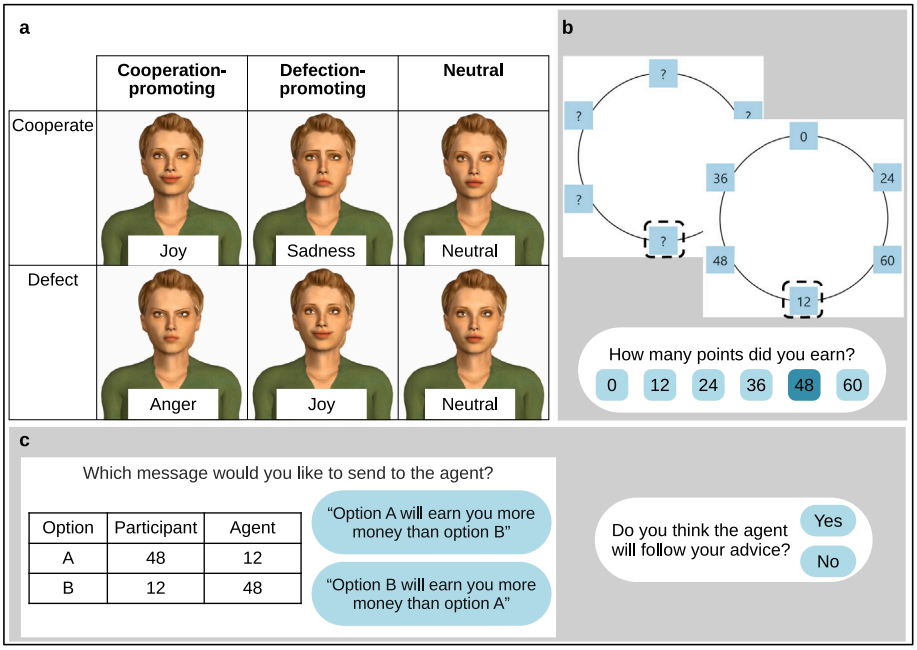


**Fig. 2. Experimental design and embodied virtual agent's emotional reactions. a:** Emotional facial reactions displayed by the agent after each participant's decision in the Prisoner's Dilemma, depending on the agent type (*cooperation-promoting, defection-promoting, neutral*). **b:** Sequence of one trial in the 20-round Mind Game. Participants mentally selected one of the six branches of the wheel (indicated by the dotted outline), observed the corresponding outcome (e.g., 12), and then (mis)reported it. The value 48 is shown in darker blue to illustrate a dishonest report. **c:** Sequence of the one-round Deception Game. Participants viewed a payoff table (left) displaying the outcomes for themselves and the agent. They then chose whether to send a truthful message (i.e., stating that Option B would give the agent higher payoff) or a deceptive one (i.e., falsely claiming that Option A would give the agent higher payoff), followed by an incentivised prediction of whether the agent would follow their advice (right). **Note.** All outcomes were expressed in points, with 12 points corresponding to €1.

Participants were randomly assigned to one of the three experimental groups 122 interacted with the *defection-promoting* agent, and 125 interacted with the *cooperation-promoting* and *neutral* agents (with 50 males in each group), and all groups completed the tasks in the same fixed order. These tasks were as follows: first, participants were asked to complete demographic questions and the Social Value Orientation scale. Then, they played the Prisoner's Dilemma Game, the Mind Game, and the Deception Game. Finally, they answered questions about their self-reported attitudes and beliefs regarding AI (all measured on a 0–1 continuous slider). The task order was fixed to prevent the more explicit honesty measure in the Deception Game from affecting choices in the Mind Game, which relies on ambiguity and self-deception (Fischbacher & Föllmi-Heusi, 2013). Before each game, participants were asked to complete comprehension checks to ensure understanding of the instructions and payoff structure. They could not proceed until they had answered all questions correctly.

### 4.5. Incentivisation procedure

Participants earned points based on their decisions in each game, with 12 points corresponding to €1. At the end of the experiment, one of the three games (the Prisoner's Dilemma, the Mind Game, or the Deception Game) was randomly selected for a bonus payment. Participants were reminded of this before and after each game. In the Prisoner's Dilemma, participants accumulated points over the 20 rounds depending on the outcomes of each decision. In the Mind Game, one out of the 20 trials was randomly selected for payment. In the Deception Game, a bonus payment (12 or 48 points) depended on whether the agent followed the participants' advice at a fixed probability of 78% across conditions, following Gneezy (2005). Participants also predicted the agent's response. Correct predictions were rewarded with 12 points to incentivise participants to provide advice that reflected their true preferences.

### 4.6. Statistical analysis

We scaled all continuous predictors (mean = 0, SD = 1). For each dependent variable, we fitted a Bayesian regression model that included the agent type, the social preferences (SVO), and, for honesty games only, the frequency of cooperation in the Prisoner's Dilemma, along with their interactions. Additional covariates (AI-related beliefs, decision-related feelings, and gender) were included as controls if not correlated (see Supplementary Materials). We specified weakly informative priors for fixed effects using normal distributions (mean = 0, SD = 0.5).

Inference was based on posterior means and 95% credible intervals (CIs). We considered an effect credible when the 95% CI excluded zero. Bayes Factors (BFs) are reported as complementary information to quantify relative evidence in favour of the alternative hypothesis versus the null hypothesis ($BF_{10}$ and $BF_{01}$, respectively). According to Wagenmakers et al. (2018), BFs between 3 and 10 are considered substantial evidence, between 10 and 100 strong evidence, and above 100 decisive evidence in favour of the corresponding hypothesis. All analyses were conducted in R (version 4.3.1; R Core Team (2018)).

Full model specifications, convergence checks and diagnostics are reported in the Supplementary Materials.

## CRediT authorship contribution statement

**Alice Cartaud:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Laurence Chaby:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Conceptualization. **Florian Pecune:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Catherine Pelachaud:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

## Data, code, and material availability

All data, analysis scripts, and experimental materials are available on OSF: (https://osf.io/9yc2v).

The preregistration, experimental code, and all R scripts used for data processing and analysis are included in the repository.

## Funding statement

This work was supported by the France 2030 grant ANR-22-PESN-0009.

Alice Cartaud's current contribution is funded by the French State through the France 2030 program and the Initiative of Excellence of the University of Lille (R-CDP-24-006-DePERU).

## Declaration of competing interest

The authors declare no competing interests.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.chbah.2026.100293.

## References

Callaway, F., Jain, Y. R., van Opheusden, B., Das, P., Iwama, G., Gul, S., Krueger, P. M., Becker, F., Griffiths, T. L., & Lieder, F. (2022). Leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences*, *119*(12), Article e2117432119.

Camerer, C. F., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: a guide for social scientists. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis (Eds.), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* (pp. 55–95). Oxford: Oxford University Press, chapter 97.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Chen, D. L., Schonger, M., & Wickens, C. (2016). Otree-an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Crivelli, C., & Fridlund, A. J. (2018). Facial displays are tools for social influence. *Trends in Cognitive Sciences*, *22*(5), 388–399.

Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., et al. (2021). The rise of affectivism. *Nature Human Behaviour*, *5*(7), 816–820.

Dupuy, L., Morin, C. M., de Sevin, E., Taillard, J., Salles, N., Bioulac, S., Auriacombe, M., Micoulaud-Franchi, J.-A., & Philip, P. (2022). Smartphone-based virtual agents and insomnia management: A proof-of-concept study for new methods of autonomous screening and management of insomnia symptoms in the general population. *Journal of Sleep Research*, *31*(2), Article e13489.

Eckel, C. C., & Wilson, R. K. (2003). The human face of game theory: Trust and reciprocity in sequential games. In *Trust and reciprocity: interdisciplinary lessons from experimental research* (pp. 245–274).

European Union (2024). Regulation (EU) 2024/1689 of the European parliament and of the council of 13 march 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). *Official Journal of the European Union*, *123*, 1–154.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Galeotti, F., Saucet, C., & Villeval, M. C. (2020). Unethical amnesia responds more to instrumental than to hedonic motives. *Proceedings of the National Academy of Sciences*, *117*(41), 25423–25428.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394.

Gneezy, U., Imas, A., & Madarász, K. (2014). Conscience accounting: Emotion dynamics and social behavior. *Management Science*, *60*(11), 2645–2658.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, *108*(2).

Isler, O., & Gächter, S. (2022). Conforming with peers in honesty and cooperation. *Journal of Economic Behavior and Organization*, *195*, 75–86.

Ito, R., De Melo, C. M., Gratch, J., & Terada, K. (2024). Emotional expression help regulate the appropriate level of cooperation with agents. In *International conference on affective computing and intelligent interaction*.

Jia, D., Romić, I., Shi, L., Su, Q., Liu, C., Liu, J., Holme, P., Li, X., & Wang, Z. (2025). Social networking agency and prosociality are inextricably linked in economic games. *Nature Human Behaviour*, *9*(12), 2620–2631.

Lanzetta, J. T., & Englis, B. G. (1989). Expectations of cooperation and competition and their effects on observers' vicarious emotional responses. *Journal of Personality and Social Psychology*, *56*(4), 543.

Lugrin, B., Pelachaud, C., & Traum, D. (2022). *The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: interactivity, platforms, application*. ACM.

de Melo, C. M., Gratch, J., Marsella, S., & Pelachaud, C. (2023). Social functions of machine emotional expressions. *Proceedings of the IEEE*, *111*(10), 1382–1397.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, *122*(22), Article e2415898122.

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631–647.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 1281–1302.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, *20*(4), 523–528.

Saga, T., Galland, L., Younsi, N., & Pelachaud, C. (2025). Greta 2.0: Social interactive agent system, optimized for neural network integration. In *Proceedings of the 25th ACM international conference on intelligent virtual agents* (pp. 1–10).

Scharlemann, J. P., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, *22*(5), 617–640.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, *44*(4), 695–729.

Shin, M., Kim, J., Van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, *120*(12), Article e2214840120.

Van Doorn, E. A., Heerdink, M. W., & Van Kleef, G. A. (2012). Emotion and the construal of social situations: Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment. *Cognition & Emotion*, *26*(3), 442–461.

Van Kleef, G. A., & Côté, S. (2022). The social effects of emotions. *Annual Review of Psychology*, *73*(1), 629–658.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., et al. (2018). Bayesian inference for psychology. part i: theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

West, C., & Zhong, C.-B. (2015). Moral cleansing. *Current Opinion in Psychology*, *6*, 221–225.

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, *8*(10), 1839–1850.