Dynamic Effects of Multi-Round Positive and Negative Feedback from Robots and Humans on Trust and Cooperation*

Hongzhou Xuan^{1,2,3}, Jiaying Zhu², Qianen Lai², Guibing He²

2026-07-17

ChinaRxiv

Abstract

Human–robot collaboration is gradually becoming an emerging mode of work in intelligent organizations. Trust is the foundation of collaboration among team members, established and dynamically evolving through interaction. Understanding the impact of human–robot interaction on trust and how it differs from interpersonal interaction is therefore particularly important. Through one pre-experiment and two formal experiments, this study examined from a dynamic perspective the effects of multi-round feedback from robot and human sources on the evolution of trust and cooperation. The pre-experiment screened appropriate positive and negative feedback statements. In the formal experiments, participants collaborated with either a robot or a human teammate to complete five rounds of investment decision-making tasks and evaluated each other between rounds. Participants received four instances of negative feedback (Experiment 1) or positive feedback (Experiment 2) from either a robot or a human teammate. The results showed that, in the context of multi-round negative feedback, trust and cooperation exhibited clearer source-related differences: under negative feedback from a robot, trust and cooperation tended to show a "delayed cumulative" evolutionary pattern, characterized by delayed responses and cumulative effects; under negative feedback from a human, trust and cooperation tended to show an "immediate response" evolutionary pattern, characterized by higher sensitivity and stronger adaptability. However, in the context of multi-round positive feedback, there were no significant differences in the dynamic effects of different feedback sources on trust and cooperation. This study provides empirical evidence from a dynamic perspective for understanding the differentiated evolution of human–robot trust and interpersonal trust, and offers implications for the design of feedback strategies in hybrid teams.

Full Text

Dynamic Effects of Multi-Round Positive and Negative Feedback from Robots and Humans on Trust and Cooperation*

Hongzhou Xuan1,2,3, Jiaying Zhu2, Qianen Lai2, Guibing He2

(1School of Business Administration, Zhejiang Gongshang University, Hangzhou 310018)
(2Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310058)
(3Business Technology Application Innovation Center, Zhejiang Gongshang University, Hangzhou 310018)

Abstract Human–machine collaboration is gradually becoming an emerging form of work in intelligent organizations. Trust is the foundation of collaborative cooperation among team members; it is established through interaction

and evolves dynamically. Understanding the influence of human–machine interaction on trust, as well as how it differs from interpersonal interaction, is therefore particularly important. Through one pre-experiment and two formal experiments, this study examined, from a dynamic perspective, the effects of multi-round feedback from robots and humans on the evolution of trust and cooperation. The pre-experiment selected appropriate positive and negative feedback statements. In the formal experiments, participants collaborated with either a robot or a human teammate to complete a five-round investment decision-making task and evaluated one another between rounds. Participants received four instances of negative feedback (Experiment 1) or positive feedback (Experiment 2) from either a robot or a human teammate. The study found that, in the context of multi-round negative feedback, trust and cooperation showed a clearer trend of source-based differences: under negative feedback from robots, trust and cooperation were more inclined to display a "lagged cumulative"pattern of evolution, characterized by delayed responses and cumulative effects; under negative feedback from humans, trust and cooperation were more inclined to display an "immediate response" pattern of evolution, characterized by higher sensitivity and stronger adaptability. However, in the context of multi-round positive feedback, there were no significant differences in the dynamic effects of different feedback sources on trust and cooperation. From a dynamic perspective, this study provides empirical evidence for understanding the differentiated evolution of human–machine trust and interpersonal trust, and offers a reference for designing feedback strategies in hybrid teams.

Keywords: human–machine collaboration; human–machine trust; evaluative feedback; trust dynamics; feedback valence

Classification number: B842

Received: 2025-05-10
*Supported by the Science and Technology Innovation 2030— "Brain Science and Brain-Inspired Research" Major Project (2021ZD0200409) and the National Natural Science Foundation of China (72571243).
Corresponding author: Guibing He, E-mail: gbhe@zju.edu.cn

1 Introduction

1.1 Human-Machine Collaboration as an Emerging Mode of Work

With the development of intelligent technologies, human society is entering the era of intelligence by leaps and bounds (Hecklau et al., 2016; Pereira et al., 2023), and people' s modes of production and daily life are being reshaped (Aazam et al., 2018; Xu et al., 2018). The social interaction relationship between human workers and robots is likewise undergoing fundamental change (Lee, 2018; Von Krogh, 2018). Unlike earlier technologies such as computers and robotic arms, which were mainly controlled by humans, intelligent robots possess cer-

tain capacities for perception, decision-making, and execution, and thus are increasingly regarded as human colleagues or collaborators (Christoffersen & Woods, 2002; Haenlein & Kaplan, 2021; Madhavan & Wiegmann, 2007; Wynne & Lyons, 2018). Similar to human members, intelligent robots can also make decisions independently, issue work instructions, evaluate performance, and provide feedback to other team members (Lv et al., 2024; Ososky et al., 2013). At present, human–machine collaborative teams have begun to be applied in fields such as healthcare, aerospace, manufacturing, and services (Deepa et al., 2024; Malik et al., 2023; Narzary et al., 2024; O' Neill et al., 2022), and the scope of their application in organizational management is also continuously expanding (Skubis & Wodarski, 2023). For example, the Polish beverage company Dictador appointed the robot Mika as chief executive officer, responsible for employee management and community communication; the robot Sophia from Hanson was assigned human-resources functions, conducting employee recruitment and performance feedback. This transformation in the mode of work may have profound effects on human employees' work behaviors, emotional experiences, and even physical and mental health (Candon et al., 2023; Li et al., 2024). Therefore, it is particularly necessary to conduct in-depth research on human–machine collaboration as a new type of human–machine relationship.

1.2 Trust and Cooperation as the Foundation of Human–Machine Collaboration

Trust is crucial to team effectiveness and team cohesion (DeRosa et al., 2004; Jones & George, 1998; Marks et al., 2001; Nicholson et al., 2001). It exists not only in interpersonal relationships (Mayer et al., 1995; Ozaras & Abaan, 2018), but is also equally critical in human–machine interaction (Gkinko & Elbanna, 2023; Komiak & Benbasat, 2006; Vance et al., 2008; Qi Yue et al., 2024).

Collaborative trust is an integrative psychological judgment formed by individuals in team collaboration (Getha-Taylor et al., 2019; Kozuch & Sienkiewicz-Małyjurek, 2022). It includes both the behavioral tendency to take action in the current task based on a collaborator' s suggestions (Lewis & Weigert, 1985; McAllister, 1995) and the willingness to establish and maintain a long-term relationship with the collaborator (Lee et al., 2011; Whitehair & Berdanier, 2017). Accordingly, this study examines collaborative trust from two aspects—behavioral trust and willingness for future cooperation. The former reflects individuals' objective reliance on teammates' suggestions in actual decision-making, while the latter reflects individuals' subjective expectations regarding future cooperative relationships.

Numerous studies have pointed out that trust in intelligent technologies is a prerequisite for individuals' willingness to use and collaborate with them (Glikson & Woolley, 2020; Habbal et al., 2024; Vanneste & Puranam, 2024). At present, although intelligent technologies are developing rapidly, their trustworthiness remains widely questioned (De Freitas et al., 2023), and insufficient trust has become a major obstacle to their broad application (Frey & Osborne, 2017;

Raisch & Krakowski, 2021). Against this background, understanding the formation and evolution of human–machine trust is of great significance for promoting efficient human–machine collaboration (De Visser et al., 2020).

1.3 Research Gaps in the Dynamics of Human–Machine Trust

Trust is formed through the accumulation of evidence and has clear dynamic characteristics (Cho et al., 2015; Kraus et al., 2020). Indi-

individuals' trust attitudes and trust behaviors dynamically evolve during interaction (Dzindolet et al., 2002; Hoff & Bashir, 2015; Roesler et al., 2024), and are continuously adjusted as the counterpart' s behavior changes (Korsgaard et al., 2018; Luo et al., 2022).

However, most existing research on human–machine trust uses single-time-point measurements (Aliasghari et al., 2021; De Visser et al., 2017; Robinette et al., 2017; Sanders et al., 2019), and mainly examines influencing factors within the "human–robot–environment" framework (Hancock et al., 2011; He Guibing et al., 2022). This static "snapshot" perspective makes it difficult to reveal the trajectory of trust change during continuous interaction (Lanz et al., 2024; Yang et al., 2023), and it also cannot capture the differences between initial trust and trust after multiple rounds of interaction (De Visser et al., 2020; Grossman & Feitosa, 2018). Therefore, shifting from a static perspective to a dynamic perspective is crucial for research on human–machine trust (Chi & Malle, 2023; Yang et al., 2023).

1.4 Feedback Interaction as a Triggering Mechanism for the Dynamic Evolution of Trust

Evaluative feedback is an important component of team interaction and also a key mechanism that triggers changes in trust. Studies show that evaluative feedback has a critical influence on trust and cooperation among members (Chen et al., 2024; Fousiani, 2020; Podsakoff et al., 2006), and employees' productivity, team relationships, and work motivation are influenced to a large extent by praise and critical feedback from colleagues or supervisors (Ali et al., 2014; Kuhnen & Tymula, 2012). As the social communication capabilities of intelligent robots improve, the importance of feedback interaction in human–machine collaborative teams is likewise becoming increasingly evident (Candon et al., 2023).

However, empirical research on how robot feedback affects the trust and cooperation of human members remains relatively limited. Our team' s previous research found that, compared with negative feedback from human sources, individuals showed higher acceptance of negative feedback from robot sources, and correspondingly exhibited higher behavioral trust and willingness to cooperate in the future (Xuan & He, 2025). Yet that study examined only the influence of

a single instance of feedback, and still could not reveal the evolutionary pattern of human–machine trust in multiple rounds of feedback interaction. Therefore, the present study further focuses on multi-round feedback contexts, examining how trust and cooperation dynamically evolve with repeated feedback, and whether the source of feedback affects this evolutionary process.

1.5 The Specificity of Robot-Sourced Feedback: Differences in Attribution and Perceived Intentionality

When people face intelligent agents and humans, they do not necessarily adopt exactly the same social strategies; this is known as the "media inequality" effect (Mou & Xu, 2017). In the domain of human–machine trust, people likewise display distinctive trust models toward intelligent agents (Hoff & Bashir, 2015; Lee & See, 2004; Lewandowsky et al., 2000). For example, when immoral behavior occurs—such as deception or discrimination—people' s punitive responses toward intelligent agents are not as strong as those toward humans (Alarcon et al., 2024; Xu Liying et al., 2022). This difference stems to a large extent from people' s different perceptions of robots and humans in terms of attribution and intentionality.

Evaluative feedback is usually understood as the feedback provider' s subjective judgment of the recipient. An individual' s response to feedback depends largely on how they interpret the feedback provider' s intention, and this process has a particularly profound impact on trust (Desmet et al., 2011; Dirks et al., 2011). There are major differences in people' s attributions and perceptions of intentionality regarding robot and human behavior. Because robots are often considered to lack autonomy and intentionality (Bigman & Gray, 2018; Gray et al., 2007), individuals are more inclined to attribute robot feedback to program settings rather than autonomous behavior (Yalcin et al., 2022), thereby potentially weakening the dynamic impact of feedback on trust. By contrast, in interpersonal interaction, others' words and behaviors are usually regarded as external reflections of their personal inner thoughts or emotions (Raver et al., 2012), and therefore interpersonal feedback may have a more pronounced

influence. On this basis, the present study hypothesizes that trust and cooperation under the influence of feedback from robot versus human sources may exhibit different patterns of dynamic evolution.

1.6 Asymmetric Effects of Positive and Negative Feedback

Feedback valence is one of the most critical characteristics of feedback (Djamasbi & Loiacono, 2008; Watts, 2007) and has a profound impact on feedback acceptance (Anseel & Lievens, 2006; Tonidandel et al., 2002). Positive feedback typically elicits positive emotions and is therefore more readily accepted (Higgins et al., 2001), whereas negative feedback may threaten self-esteem (Malle et al., 2014) and trigger strong negative emotions and adverse reactions (Weiss et al., 1999). Thus, positive and negative feedback may have different effects on

trust and cooperation.

Research on information processing shows that individuals generally exhibit a "negativity bias," that is, they allocate more attention and cognitive resources to negative information (Baumeister et al., 2001; Carretié et al., 2001; Ito et al., 1998; Rozin & Royzman, 2001). In the process of trust formation, negative events also often produce stronger and more lasting effects than positive events (Yang et al., 2023). Our team' s previous research (Xuan & He, 2025) also revealed a significant moderating role of feedback valence, finding that the effect of human–machine feedback source was significant only under negative feedback conditions. We argue that, as information that may threaten self-esteem, negative feedback is more likely to trigger individuals' defense mechanisms and attributional processes; therefore, under such conditions, differences in feedback-source effects are more likely to be amplified and attended to. By contrast, as self-enhancing information, positive feedback is usually accepted directly by individuals, with less inference about the source' s intentions or deep processing. Accordingly, the present study hypothesizes that the effects of multi-round positive and negative feedback on trust and cooperation may be asymmetric. Compared with positive feedback, negative feedback may induce more pronounced dynamic changes, and differences between robot and human feedback sources may also be more likely to emerge in negative-feedback contexts.

1.7 Research Overview

Through one pre-experiment and two formal experiments, the present study examines the influence of multi-round feedback from robot and human sources on the dynamic evolution of trust and cooperation. The pre-experiment was used to screen positive and negative feedback materials with clear valence and comparable tone intensity. Experiment 1 examined multi-round negative feedback from robot and human sources, and Experiment 2 examined multi-round positive feedback from the two sources. In the two formal experiments, participants completed a five-round cooperative investment task with either a robot or a human teammate and received four instances of evaluative feedback between rounds. The study measured behavioral trust and willingness to cooperate in the future through advice-adoption behavior and subjective scales, respectively, with a focus on examining stage-by-stage changes in trust and cooperation during the continuous feedback process as well as differences in their trajectories.

2 Pre-experiment: Screening of Feedback Statement Materials

2.1 Experimental Purpose

Before the formal experiments, a pre-experiment needed to be conducted to identify appropriate feedback statement materials. Although previous studies have provided many feedback statements, these materials have problems such as poor fit with the scenario, inconsistent format, unclear valence, and large

differences in tone intensity, and therefore cannot be directly applied to the present study. Accordingly, this pre-experiment aimed to screen positive and negative feedback statements suitable for the subsequent formal experimental context and to ensure that these statements remained relatively consistent in format and tone intensity.

2.2 Experimental Method

2.2.1 Participants Participants in the pre-experiment were recruited through the sample pool of the online survey platform Wenjuanxing. A total of 200 people participated in this experiment. Exclu-

After excluding 8 participants who withdrew midway or selected the same option throughout, the final valid sample size was 192, including 97 men, with a mean age of 30.48 years (SD = 6.79 years). All participants received 2 RMB as compensation after completing the experiment.

2.2.2 Experimental Procedure

Following the material-screening procedure of Hicks et al. (2020), this pilot experiment was divided into three stages: material collection, preliminary screening and revision, and material rating.

(1) Material collection. Using five large language models, including Tongyi Qianwen, Doubao, and Kimi, and taking the scenarios and experimental design of the subsequent experiment as prompts, each model generated 4 negative feedback statements and 4 positive feedback statements, yielding a total of 40 feedback statements.

(2) Preliminary screening and revision. Two PhDs in psychology conducted preliminary screening and revision of the 40 feedback statements, deleting statements that were too long, too short, or overly exaggerated in tone, and revising expressions that were overly formal or written. At this stage, 20 feedback statements were ultimately retained, including 10 negative feedback statements and 10 positive feedback statements.

(3) Material rating. The 192 participants rated the tone intensity of the 20 feedback statements (1 = "very weak," 7 = "very strong"). All feedback statements were presented in random order.

Table 1 Tone-intensity scores of feedback statements in the pilot experiment

Negative feedback	Mean	Standard deviation	Positive feedback	Mean	Standard deviation
N-1	3.84	1.59	P-1	4.81	1.49
N-2	3.93	1.44	P-2	4.70	1.37
N-3	4.35	1.49	P-3	4.48	1.40

Negative feedback	Mean	Standard deviation	Positive feedback	Mean	Standard deviation
N-4	4.49	1.45	P-4	5.19	1.46
N-5	5.41	1.21	P-5	5.46	1.38
N-6	5.16	1.33	P-6	5.10	1.29
N-7	4.83	1.40	P-7	5.04	1.38
N-8	5.00	1.42	P-8	5.42	1.34
N-9	5.18	1.39	P-9	5.76	1.42
N-10	5.38	1.50	P-10	5.82	1.34

2.2.3 Experimental Results

The tone-intensity ratings of the 20 feedback statements are shown in Table 1; higher scores indicate a stronger tone. Because an excessively strong tone would lead participants to feel that the situation was artificial, thereby casting doubt on the authenticity of the teammate' s identity and the cooperative relationship, this experiment selected the statements with scores closest to 5. This both ensured a tone effect that was moderately above average and reduced participants' doubts about the authenticity of the research situation. Ultimately, the negative feedback statements N-6, N-7, N-8, and N-9, and the positive feedback statements P-1, P-4, P-6, and P-7 were retained.

One-sample t tests showed that the tone intensity of the above eight selected statements did not significantly deviate from the criterion value of 5 ($ps > 0.071$). A paired-samples t test further showed that there was no significant difference between the overall tone intensity of negative feedback ($M = 5.04$, $SD = 1.03$) and that of positive feedback ($M = 5.03$, $SD = 1.10$), $t(191) = 0.08$, $p = 0.937$. In summary, the screening of feedback statements achieved the expected effect; the specific content is shown in Table 2.

Table 2. Feedback statements finally confirmed in the pre-experiment

Negative feedback	Negative feedback	Positive feedback	Positive feedback
No.	Content	No.	Content
N-6	This result is somewhat disappointing. You may not have been focused enough; we will have to work harder if we are to make it into the top five.	P-1	This result is already very good. I believe you still have even greater potential. If we work hard, we can make it into the top five!

Negative feedback	Negative feedback	Positive feedback	Positive feedback
N-7	Our performance has not improved much. I think your decisions may have been too casual and, overall, not meticulous enough.	P-4	I think you have worked very hard, and throughout the process I could also see you actively making adjustments. Keep going—make one final push.
N-8	We do not have many chances left. Your attitude needs to be a bit more proper and serious; reflect on the mistakes that may have existed before.	P-6	Our scores are still relatively close. I believe that as long as you are a little more meticulous, our performance will certainly reach a higher level.
N-9	I think that, overall, you still have not worked hard enough; the adjustments during the process were also not sufficiently in place, and your decisions were too casual and not meticulous enough.	P-7	We still have a chance. As long as we maintain a proper and serious attitude and make slight adjustments to the mistakes that may have existed before.

3 Experiment 1: Dynamic Effects of Multiple Rounds of Negative Feedback From Robot and Human Sources on Trust and Cooperation

3.1 Experimental Objective

Experiment 1 aimed to explore the differential effects of multiple rounds of negative feedback from robot and human teammates on the dynamic evolution of trust and cooperation. In this experiment, either a robot or a human teammate

provided participants with four consecutive rounds of negative feedback.

3.2 Experimental Method

3.2.1 Participants Referring to the power-parameter settings adopted in similar human–robot trust studies (Roesler et al., 2024), an a priori power analysis was conducted for this experiment. For a 2×5 mixed-design repeated-measures analysis of variance, to achieve a medium effect size (0.25) at a significance level of $\alpha = 0.05$ and statistical power close to 0.90 $(1 - \beta)$, the minimum required sample size was 28 participants. To meet the minimum sample-size requirement and to account for possible participant exclusions caused by technical or system problems, the target sample size for this experiment was set at 56 participants.

Using convenience sampling, participants were recruited from universities in an eastern region through campus forums and social media platforms. A total of 56 participants were recruited for Experiment 1, including 18 males, with a mean age of 21.34 years ($SD = 2.61$ years). Participants were randomly assigned to either the robot group or the human group, with 28 participants in each group. All participants were physically and mentally healthy, had no cognitive impairments, and came from majors unrelated to psychology or robotics; none had prior experience of close interaction with robots. Each participant signed an informed-consent form before the experiment and received RMB 40 as compensation after the experiment.

Figure 1. Image of the Pepper robot and experimental interaction scenario

3.2.2 Experimental Design

Experiment 1 adopted a 2×5 mixed experimental design. The between-subjects variable was feedback source (robot teammate or human teammate), and the within-subjects variable was interaction phase (before feedback, T0; after the first feedback, T1; after the second feedback, T2; after the third feedback, T3; and after the fourth feedback, T4). Four instances of negative feedback were administered during the experiment.

The dependent variables included behavioral trust and willingness for future cooperation. Behavioral trust was measured using an adapted TNO trust paradigm, assessing the objective degree to which participants relied on their teammate' s recommendation in actual decision-making. Willingness for future cooperation was measured with a scale assessing participants' subjective willingness to continue cooperating with the teammate.

3.2.3 Experimental Materials

The experiment used an investment decision-making task developed on the basis of real fund indicators. Participants were required to judge their willingness to invest according to indicators such as a fund' s historical return rate and portfolio structure. The experiment presented only preset team rankings and

teammate feedback, and did not provide verifiable information about individual performance, in order to reduce interference from objective performance on the feedback effect.

In the robot condition, the humanoid social robot Pepper, developed by SoftBank Robotics, served as the teammate (see Figure 1). Pepper' s speech, gaze, and arm movements were controlled using Choregraphe 2.1.4 software. In the human condition, the teammate was played by a human actor of the same gender as the participant. The actor interacted with the participant according to a standardized script, with the content of expression and the timing of interaction kept as consistent as possible with those in the robot condition.

The measurement of behavioral trust was adapted from the TNO trust paradigm and focused on the degree of reliance on the teammate' s recommendation in actual decision-making (De Visser et al., 2016, 2020; Kulms & Kopp, 2019; Van Dongen & Van Maanen, 2013). In each trial, participants first reported an initial answer and then, after receiving the teammate' s recommendation, confirmed a final answer. Behavioral trust was calculated according to the extent to which the final answer was adjusted from the initial answer toward the teammate' s recommendation:

$$\text{Behavioral trust} = \frac{\text{final answer} - \text{initial answer}}{\text{teammate' s recommendation} - \text{initial answer}}$$

The theoretical range of this index is $[0, 1]$; higher values indicate that participants were more inclined to adopt the teammate' s recommendation, that is, that their level of behavioral trust in the teammate was higher. Following previous procedures, if behavioral trust exceeded the theoretical interval, the data for that trial needed to be winsorized (Lanz et al., 2024; Logg et al., 2019). However, in Experiment 1, no values outside the theoretical interval were found

unreasonable extreme values. Only 1.93% of trials (27 trials, involving 14 participants) were excluded because the teammate' s suggestion was exactly the same as the participant' s initial answer (i.e., the denominator of the formula was 0).

The future willingness-to-collaborate scale was adapted from previous studies (Gross et al., 2018; Van Pinxteren et al., 2019) and contained two items: "If there is a similar task, I would be willing to continue collaborating with this teammate" and "In other tasks, I would be willing to continue collaborating with this teammate." Participants rated the items on a 7-point Likert scale (1 = "very unwilling," 7 = "very willing"). In the present experiment, Cronbach' s α for this scale was 0.95, indicating good reliability.

Figure 2. Overall experimental procedure of Experiment 1

Visible labels in the figure:

- 3 trials

- 5 trials
- Feedback 1
- Feedback 2
- Feedback 3
- Feedback 4
- T0
- T1
- T2
- T3
- T4
- Baseline group
- Experimental group
- Independent investment task
- Collaborative investment task
- Future willingness-to-collaborate scale
- Feedback acceptance scale

Figure 3. System interfaces for the independent investment task and the collaborative investment task

Independent investment task

Which of the following two funds would you be more willing to invest in?

A: Zhongcheng Growth Fund

Historical performance:

- Average annual return: 8%
- Volatility: 10%

Risk-adjusted return:

- Sharpe ratio: 0.8
- Information ratio: 0.7

Fund fees:

- Management fee: 1.2%
- Subscription/redemption fee: no subscription fee; redemption fee 0.5%

Fund team:

- Fund manager: 10 years of experience; has managed this fund for 5 years
- Investment team: stable team; core members have been in their positions for more than 5 years

Portfolio structure:

- 60% stocks, 30% bonds, 10% foreign exchange

B: Linghang Technology Fund

Historical performance:

- Average annual return: 12%
- Volatility: 15%

Risk-adjusted return:

- Sharpe ratio: 0.7
- Information ratio: 0.8

Fund fees:

- Management fee: 1.5%
- Subscription/redemption fee: subscription fee 1%; no redemption fee

Fund team:

- Fund manager: 5 years of experience; has managed this fund for 3 years
- Investment team: stable team; members have strong professional backgrounds

Portfolio structure:

- 80% stocks, 10% real estate, 10% flexible assets

Collaborative investment task

What is your investment intention toward the following fund?

Ancheng Mixed Fund

Historical performance:

- Average annual return: 7%
- Volatility: 8%

Risk-adjusted return:

- Sharpe ratio: 0.9
- Information ratio: 0.6

Fund fees:

- Management fee: 1.1%
- Subscription/redemption fee: no subscription fee; redemption fee 0.2%

Fund team:

- Fund manager: 10 years of experience; has managed this fund for 5 years
- Investment team: stable team; members jointly manage multiple funds

Investment strategy and portfolio structure:

- Investment objective and strategy: conservative returns; mainly invests in bonds and large-cap blue-chip stocks

- Portfolio structure: 40% stocks, 40% bonds, 10% cash, 10% bank wealth-management products

3.2.4 Experimental Procedure

The overall experimental procedure is shown in Figure 2. After arriving at the laboratory and signing the informed-consent form, participants were randomly assigned to either the robot group or the human group. Participants in the robot group first engaged in a brief interaction with Pepper; participants in the human group met an actor posing as another participant and gave a simple self-introduction. Participants then read the experimental instructions and learned about the collaborative task, the scoring method, and the fact that teams ranked in the top five would receive additional rewards.

The formal experiment consisted of two phases: independent investment and collaborative investment. The system interface is shown in Figure 3. In the independent-investment phase, participants and their teammates separately completed three fund-selection trials; that is, based on the key indicators provided, they chose the one they were more willing to invest in from two funds of the same type. Participants were told that the funds' actual returns from the previous week would be used as the criterion, and that the higher the return of the fund they selected, the higher their score would be. The task was relatively difficult: participants could not know their teammate' s score and also found it difficult to estimate their own actual performance. After this phase ended, the system displayed a preset, fictitious team ranking and individual score: "Team ranking: 6th; you outperformed 68% of previous participants; your teammate outperformed 77% of previous participants."

In the collaborative-investment phase, participants completed five rounds of tasks (T0–T4), with each round containing five trials. In each trial, the system presented detailed information about one fund, and the participant and teammate were required to evaluate their investment intention for that fund ($0 \sim 100$). Before the first round began, participants were assigned the role of team leader through a fictitious lottery mini-game, meaning that they had the final decision-making authority. Therefore, in each trial, the participant first made an initial judgment about their investment intention for the fund. Pepper or the actor then immediately received a random value generated by the system (within a range of ± 30 from the initial judgment). After pretending to think briefly, they orally reported this value as their recommendation and pretended to enter it into the system (Pepper was disguised as transmitting the value into the system via a data cable). The participant then had to confirm the final answer again. After each round of tasks, participants completed a future willingness-to-cooperate scale.

Between the five rounds of collaborative-investment tasks, participants were required to conduct four face-to-face oral evaluations with their teammate. Participants first evaluated their teammate, and then the teammate directly provided

oral feedback according to the scripted content. In this experiment, the teammate gave negative feedback in all four consecutive instances, using four negative statements determined in the pre-experiment. After all tasks were completed, participants filled out a demographic questionnaire.

3.3 Experimental Results

3.3.1 Dynamic Evolution of Trust and Cooperation This experiment used a 2 (feedback source) × 5 (interaction stage) mixed analysis of variance (mixed ANOVA) for statistical testing, with interaction stage as the repeated-measures factor. Considering that the probability of adopting advice might show a nonlinear change with the magnitude of the difference (|teammate recommendation − initial answer|), we included it as a covariate. Simple-effect comparisons were corrected for multiple comparisons using the Holm method ($\alpha = 0.05$).

Y-axis: Behavioral trust
X-axis: Interaction stage
Legend: Feedback source –Human; Robot

Figure 4. Dynamic evolution of behavioral trust under multiple rounds of negative feedback

Note: Markers indicate means; error bands indicate standard errors.

The changes in behavioral trust are shown in Figure 4. The results indicated that the main effect of feedback source was not significant ($F(1, 54) = 0.01, p = 0.947$), suggesting that, overall, participants' behavioral trust in robot teammates ($M = 0.46, SD = 0.20$) did not differ significantly from their behavioral trust in human teammates ($M = 0.45, SD = 0.20$). The main effect of interaction stage was significant ($F(4, 216) = 7.43, p < 0.001$, partial $\eta^2 = 0.12$), with behavioral trust showing an overall downward trend from stage T0 ($M = 0.54, SD = 0.15$) to stage T4 ($M = 0.41, SD = 0.23$). The interaction between feedback source and interaction stage was marginally significant ($F(4, 216) = 2.30, p = 0.060$, partial $\eta^2 = 0.04$), suggesting that the cross-stage pattern of change in behavioral trust may differ across feedback-source conditions.

Simple-effects analyses showed that, under the robot-source feedback condition, the changes in behavioral trust between adjacent stages were not significant from T0 ($M = 0.55, SD = 0.18$) to T1 ($M = 0.52, SD = 0.20$), from T1 to T2 ($M = 0.45, SD = 0.16$), from T2 to T3 ($M = 0.39, SD = 0.20$), or from T3 to T4 ($M = 0.38, SD = 0.23$) ($p_{Holms} > 0.240$). However, T4 was significantly lower than T0 ($p_{Holm} = 0.003$), showing a gradual downward trend. Under the human-source feedback condition, behavioral trust decreased significantly from T0 ($M = 0.54, SD = 0.13$) to T1 ($M = 0.44, SD = 0.20$) ($p_{Holm} = 0.045$). Thereafter, no significant changes were observed from T1 to T2 ($M = 0.41, SD = 0.19$), from T2 to T3 ($M = 0.46, SD = 0.23$), or from T3 to T4

($M = 0.43, SD = 0.24$) ($p_{Holms} > 0.242$). T4 was significantly lower than T0 ($p_{Holm} = 0.038$), indicating a pattern of sharp decline followed by stabilization.

Figure labels: y-axis, "Future cooperation willingness" ; x-axis, "Interaction stage" (T0, T1, T2, T3, T4); legend, "Feedback source" : Human, Robot.

Figure 5 Dynamic evolution of future cooperation willingness under multiple rounds of negative feedback

Note: Markers indicate means; error bands indicate standard errors.

The change in future cooperation willingness is shown in Figure 5. The results showed that the main effect of feedback source was not significant ($F(1, 54) = 1.33, p = 0.254$), indicating that, overall, participants' overall future cooperation willingness toward robot teammates ($M = 4.33, SD = 1.72$) was similar to that toward human teammates ($M = 3.88, SD = 1.89$). The main effect of interaction stage was significant ($F(4, 216) = 54.12, p < 0.001,$ partial $\eta^2 = 0.50$); future cooperation willingness declined overall from the T0 stage ($M = 5.34, SD = 1.12$) to the T4 stage ($M = 3.29, SD = 1.89$). The interaction between feedback source and interaction stage was significant ($F(4, 216) = 3.70, p = 0.006,$ partial $\eta^2 = 0.06$).

Simple-effects analyses showed that the stage-by-stage change patterns in future cooperation willingness differed across feedback-source conditions. Under the robot-source feedback condition, future cooperation willingness did not change significantly from T0 ($M = 5.43, SD = 1.17$) to T1 ($M = 5.04, SD = 1.22$) ($p_{\text{Holm}} = 0.163$), but declined significantly from T1 to T2 ($M = 4.23, SD = 1.44$) ($p_{\text{Holm}} < 0.001$), and from T2 to T3 ($M = 3.61, SD = 1.85$) ($p_{\text{Holm}} = 0.009$). It then tended to stabilize again from T3 to T4 ($M = 3.34, SD = 1.87$) ($p_{\text{Holm}} = 0.181$), but T4 was significantly lower overall than T0 ($p_{\text{Holm}} < 0.001$), showing a pattern of gradual decline. Under the human-source feedback condition, future cooperation willingness declined significantly from T0 ($M = 5.25, SD = 1.08$) to T1 ($M = 4.00, SD = 1.81$) ($p_{\text{Holm}} < 0.001$), and also declined significantly from T1 to T2 ($M = 3.43, SD = 1.86$) ($p_{\text{Holm}} = 0.002$). Thereafter, it remained stable from T2 to T3 ($M = 3.45, SD = 1.96$) and from T3 to T4 ($M = 3.25, SD = 1.95$) ($p_{\text{HolmS}} > 0.869$). Overall, it likewise showed a significant decrease from T0 to T4 ($p_{\text{Holm}} < 0.001$), displaying a pattern of rapid early decline followed by stabilization.

3.3.2 Segmented evolutionary patterns of trust and cooperation

To further characterize the trajectories of change in trust and cooperation, this study further used mixed-effects models for analysis and systematically compared three types of models: (1) a linear model, assuming that trust changes linearly over time; (2) a quadratic model, allowing trust to follow a curvilinear trajectory; and (3) a segmented model, in which, based on the results of the mixed ANOVA and the evolution-process diagrams, the interaction process was divided into two stages—an early stage (T0–T1) and a later stage (T2–T4)—

and the rate of change at each stage was estimated separately under different feedback-source conditions. All models included feedback source and interaction stage as predictors, with behavioral trust and future cooperation willingness as

as dependent variables, and both included random intercepts and random slopes at the participant level to capture individual differences. Model fitting used maximum likelihood estimation, and model comparisons were conducted using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Burnham & Anderson, 2004). All analyses were completed in R (Version 4.4.0) using the *lme4* and *lmerTest* packages. Full model information is provided in Appendix A.

The model comparison results showed that the segmented model fit relatively better for both behavioral trust (AIC = −181.33, BIC = −148.62) and future willingness to cooperate (AIC = 806.99, BIC = 839.70). This indicates that dividing the interaction process into two phases, early and late, can more accurately characterize the dynamic evolution of trust and cooperation.

The slope estimates from the segmented mixed-effects model showed that the rate of change in trust and cooperation differed significantly between the early phase (T0–T1) and the late phase (T2–T4) under different feedback-source conditions. In the robot condition, behavioral trust did not change significantly in the early phase ($\beta = -0.03$, $SE = 0.03$, $p = 0.209$; $\beta_{\text{standardized}} = -0.09$, 95% CI [−0.24, 0.05]), but declined significantly in the late phase ($\beta = -0.05$, $SE = 0.02$, $p = 0.001$; $\beta_{\text{standardized}} = -0.24$, 95% CI [−0.39, −0.10]), reflecting a delayed-response pattern. By contrast, in the human condition, behavioral trust declined sharply in the early phase ($\beta = -0.11$, $SE = 0.03$, $p < 0.001$; $\beta_{\text{standardized}} = -0.27$, 95% CI [−0.41, −0.12]), whereas in the late phase it tended to stabilize ($\beta = 0.01$, $SE = 0.02$, $p = 0.940$; $\beta_{\text{standardized}} = 0.01$, 95% CI [−0.14, 0.15]), showing an immediate-response pattern.

Future willingness to cooperate exhibited a similar pattern. The robot group declined slowly in the early phase ($\beta = -0.50$, $SE = 0.18$, $p = 0.005$; $\beta_{\text{standardized}} = -0.13$, 95% CI [−0.23, −0.04]), and the magnitude of decline increased further in the late phase ($\beta = -0.60$, $SE = 0.08$, $p < 0.001$; $\beta_{\text{standardized}} = -0.33$, 95% CI [−0.42, −0.24]), reflecting a cumulative-effect pattern. In contrast, the human group declined rapidly in the early phase ($\beta = -1.40$, $SE = 0.18$, $p < .001$; $\beta_{\text{standardized}} = -0.38$, 95% CI [−0.47, −0.29]), while the rate of decline in the late phase clearly slowed ($\beta = -0.21$, $SE = 0.08$, $p = 0.009$; $\beta_{\text{standardized}} = -0.12$, 95% CI [−0.21, −0.03]), reflecting a rapid-adaptation pattern.

3.3.3 Robustness Test of Suggestion Adoption

To test the robustness of the behavioral-trust results, the study further constructed a binary suggestion-adoption indicator based on whether the participant "crossed the midpoint between the initial answer and the teammate' s suggestion," and analyzed it using a generalized linear mixed-effects model (Gen-

eralized Linear Mixed Model, GLMM). The model included feedback source, interaction phase, and discrepancy magnitude, and set random intercepts for participants.

The results showed that the main effect of feedback source was not significant ($\chi^2(1) = 0.43$, $p = 0.512$), indicating no significant overall difference in suggestion adoption between robot teammates and human teammates; the main effect of interaction phase was significant ($\chi^2(4) = 102.14$, $p < 0.001$), indicating that suggestion adoption changed significantly over the interaction process; importantly, the interaction effect between feedback source and interaction phase was significant ($\chi^2(4) = 20.93$, $p < 0.001$), indicating significant differences in the dynamic evolution of suggestion adoption under different feedback sources.

Descriptive statistics (see Table 3) showed that the adoption rate in the robot-teammate group gradually decreased from 76.47% at T0 to 31.16% at T4, showing a stepwise declining trend. In contrast, the suggestion-adoption rate in the human-teammate group dropped rapidly from 68.12% at T0 to 45.99% at T1, and then remained at a similar level, showing a pattern of stabilizing after a sharp decline. The results of this analysis were highly consistent with the model for behavioral-trust measurement, further confirming the robustness of the main analysis results.

Table 3 Descriptive statistics of advice adoption in Experiment 1

	Robot group	Human group
T0	76.47%	68.12%
T1	68.66%	45.99%
T2	49.64%	38.85%
T3	36.76%	45.00%
T4	31.16%	39.86%

3.3.4 Overall change in trust and cooperation In this experiment, the difference between T4 and T0 (T4–T0) was used to represent the overall amount of change, and an independent-samples t test was further used for statistical analysis. The results showed that the overall change in behavioral trust after multiple rounds of negative feedback from the robot ($M = -0.17$, $SD = 0.27$) did not differ significantly from that after multiple rounds of negative feedback from humans ($M = -0.11$, $SD = 0.27$), $t = 0.791$, $p = 0.433$, Cohen' s $d = -0.21$. The overall change in future willingness to cooperate after multiple rounds of negative feedback from the robot ($M = -2.09$, $SD = 1.49$) also did not differ significantly from that after multiple rounds of negative feedback from humans ($M = -2.00$, $SD = 1.49$), $t = 0.224$, $p = 0.824$. This indicates that the final magnitude of decline caused by the two sources of feedback was similar, but the process of decline differed.

3.4 Summary of the experiment

Experiment 1 showed that, under the condition of multiple rounds of negative feedback, the source of feedback exerted a significantly differentiated influence on the dynamic evolution of trust and cooperation. Under the robot-source feedback condition, trust and cooperation changed relatively little in the initial stage but continued to decline in the later stage; under the human-source feedback condition, trust and cooperation declined significantly immediately after the first negative feedback (the rate of decline was nearly three times that under the robot condition), whereas the changes in the later stage tended to be more gradual.

Based on the above findings, we summarized the dynamic evolution of trust and cooperation under the two sources of feedback as follows: in the face of multiple rounds of negative feedback from a robot source, trust and cooperation showed a "lagged cumulative" evolutionary pattern (delayed response and cumulative effect); in the face of multiple rounds of negative feedback from a human source, trust and cooperation showed an "immediate response" evolutionary pattern (high sensitivity and strong adaptability). This finding highlights the importance of studying human–robot trust from a dynamic perspective.

4 Experiment 2: Dynamic effects of multiple rounds of positive feedback from robot and human sources on trust and cooperation

4.1 Purpose of the experiment

Experiment 2 aimed to explore the influence of multiple rounds of positive feedback from robot and human teammates on the dynamic evolution of trust and cooperation. In this experiment, the robot or human teammate provided participants with four consecutive rounds of positive feedback.

4.2 Experimental method

4.2.1 Participants Experiment 2 used the same criteria for determining sample size and the same recruitment method as Experiment 1. A total of 56 participants were recruited, including 26 males, with a mean age of 21.36 years ($SD = 2.78$ years). Participants were randomly assigned to the robot group or the human group, with 28 in each group

participants. All participants were physically and mentally healthy, had no cognitive impairments, and came from majors unrelated to psychology or robotics; none had prior experience of close interaction with robots. Each participant signed an informed-consent form before the experiment and received RMB 40 as compensation after the experiment.

4.2.2 Experimental Design Experiment 2 adopted a 2×5 mixed experimental design. The between-participants variable was feedback source (robot

teammate vs. human teammate), and the within-participants variable was interaction stage (before feedback, T0; after the first feedback, T1; after the second feedback, T2; after the third feedback, T3; after the fourth feedback, T4). Except that the feedback valence was changed to positive, the experimental design and dependent variables were the same as in Experiment 1.

4.2.3 Experimental Materials The experimental materials were the same as in Experiment 1. In Experiment 2, no unreasonable extreme values outside the theoretical range were found. Only 2.00% of trials (28 trials, involving 13 participants) were excluded because the teammate' s suggestion was exactly the same as the participant' s initial answer. In addition, Cronbach' s α for the future willingness to cooperate scale was 0.903, indicating good reliability.

4.2.4 Experimental Procedure The overall experimental procedure was the same as in Experiment 1. The only difference was that the teammate provided positive feedback four consecutive times between the five rounds of the collaborative investment task; the feedback content consisted of four positive statements selected in the pre-experiment.

4.3 Experimental Results

4.3.1 Dynamic Evolution of Trust and Cooperation Statistical testing in this experiment was conducted using a 2 (feedback source) $\times$5 (interaction stage) mixed analysis of variance, with interaction stage serving as the repeated-measures factor. Likewise, the magnitude of difference was included in the analysis as a covariate, and the Holm method ($\alpha = 0.05$) was used to correct for multiple comparisons of simple effects.

Changes in behavioral trust are shown in Figure 6. The results indicated that the main effect of feedback source was not significant ($F(1, 54) = 1.70, p = 0.198$): overall behavioral trust toward robot teammates ($M = 0.57, SD = 0.16$) did not differ significantly from overall behavioral trust toward human teammates ($M = 0.54, SD = 0.14$). The main effect of interaction stage was not significant ($F(4, 216) = 0.50, p = 0.739$): behavioral trust remained relatively stable overall from the T0 stage ($M = 0.57, SD = 0.13$) to the T4 stage ($M = 0.54, SD = 0.18$). The interaction between feedback source and interaction stage was likewise not significant ($F(4, 216) = 0.47, p = 0.761$).

Changes in future willingness to cooperate are shown in Figure 7. The results indicated that the main effect of feedback source was not significant ($F(1, 54) = 1.48, p = 0.229$): overall future willingness to cooperate with robot teammates ($M = 5.67, SD = 0.99$) did not differ significantly from overall future willingness to cooperate with human teammates ($M = 5.38, SD = 1.08$). The main effect of interaction stage was not significant ($F(4, 216) = 1.45, p = 0.218$): future willingness to cooperate remained relatively stable overall from the T0 stage ($M = 5.38, SD = 1.03$) to the T4 stage ($M = 5.47, SD = 1.18$). The interac-

tion between feedback source and interaction stage was likewise not significant ($F(4, 216) = 0.57, p = 0.686$).

Figure labels: y-axis, "Behavioral trust" ; x-axis, "Interaction stage" (T0, T1, T2, T3, T4); legend, "Feedback source" : Human, Robot.

Figure 6. Dynamic evolution of behavioral trust under multi-round positive feedback

Note: Markers indicate means, and error bands indicate standard errors.

Figure labels: y-axis, "Willingness for future cooperation" ; x-axis, "Interaction stage" (T0, T1, T2, T3, T4); legend, "Feedback source" : Human, Robot.

Figure 7. Dynamic evolution of willingness for future cooperation under multi-round positive feedback

Note: Markers indicate means, and error bands indicate standard errors.

4.3.2 Staged Evolution Patterns of Trust and Cooperation

To examine whether trust and cooperation exhibit potential dynamic patterns of change under multi-round positive feedback, this study likewise used mixed-effects models for analysis, and compared the fit performance of the linear model, quadratic model, and piecewise model. Complete model information and parameters are provided in Appendix A.

The model comparison results showed that, for the behavioral trust indicator, the linear mixed-effects model performed best (AIC = −303.05, BIC = −273.97); for the willingness for future cooperation indicator, differences in model fit were small across models, and the linear model was selected based on the principle of model parsimony (AIC = 638.16, BIC = 667.24).

The results of the linear mixed-effects model indicated that the interaction-stage effects and feedback ···

source effect and the interaction effect between the two were both nonsignificant ($ps \geq 0.063$). Therefore, under the condition of multiple rounds of positive feedback, trust and cooperation did not show stable linear changes, stagewise turning points, or source differences.

4.3.3 Robustness test for advice adoption To further verify the robustness of the above results, we conducted supplementary logistic regression analyses. A generalized linear mixed-effects model analysis using the binary advice-adoption indicator showed that the main effect of feedback source was nonsignificant ($\chi^2(1) = 0.04,\ p = 0.838$); the main effect of interaction stage was significant ($\chi^2(4) = 9.86,\ p = 0.043$); and the interaction effect between feedback source and interaction stage was nonsignificant ($\chi^2(4) = 5.61,\ p = 0.231$). Although the advice-adoption rate fluctuated to some extent across interaction stages, the change trajectories of the robot group and the human group did not

differ significantly. The descriptive statistics (see Table 4) show that, whether in the robot teammate group or the human teammate group, the advice-adoption rate remained at a relatively high level throughout the process of multiple rounds of positive feedback.

Table 4 Descriptive statistics of advice adoption in Experiment 2

	Robot group	Human group
T0	78.68%	71.43%
T1	69.78%	69.57%
T2	74.07%	67.65%
T3	68.38%	74.82%
T4	61.76%	65.94%

4.3.4 Overall change in trust and cooperation To explore the influence of the source of multiple rounds of positive feedback on the overall change in trust and cooperation, this experiment used the difference between T4 and T0 (T4−T0) to represent the overall amount of change, and further conducted statistical analyses using two-tailed independent-samples t tests. The results showed that the overall change in behavioral trust after multiple rounds of positive feedback from the robot ($M = -0.06,\ SD = 0.26$) did not differ significantly from that after human feedback ($M = -0.01,\ SD = 0.17$) ($t = 0.80,\ p = 0.427$); likewise, the overall change in future willingness to cooperate after multiple rounds of positive feedback from the robot ($M = 0.02,\ SD = 1.08$) did not differ significantly from that after human feedback ($M = 0.18,\ SD = 1.01$) ($t = 0.58,\ p = 0.567$).

4.3.5 Integrated analysis of how different feedback valences affect the dynamic evolution patterns of trust and cooperation To systematically examine the joint effects of feedback source, feedback valence, and interaction stage on trust and cooperation, this study combined the data from Experiment 1 and Experiment 2 and conducted a 2 (feedback source) ×2 (feedback valence) ×5 (interaction stage) three-factor mixed analysis of variance, in which feedback source and feedback valence were between-subjects variables, and interaction stage was the repeated-measures factor.

The analysis of behavioral trust showed that the main effect of interaction stage was significant ($F(4, 432) = 6.10,\ p < 0.001,\ partial\ \eta^2 = 0.05$), with behavioral trust decreasing significantly over the course of interaction; the main effect of feedback source was nonsignificant ($F(1, 108) = 0.81,\ p = 0.37$); the main effect of feedback valence was significant ($F(1, 108) = 18.95,\ p < 0.001,\ partial\ \eta^2 = 0.15$), and the level of behavioral trust under the negative-feedback condition ($M = 0.46,\ SD = 0.20$) was significantly lower than that under the positive-feedback condition ($M = 0.55,\ SD = 0.15$). The interaction effect between interaction stage and feedback source was marginally significant ($F(4, 432) =$

2.36, $p = 0.053$, *partial* $\eta^2 = 0.02$), indicating that the feedback source affected behav-

had a certain impact on the cross-stage pattern of change in trust; the interaction between interaction stage and feedback valence was significant ($F(4, 432) = 3.24, p = 0.012, partial\ \eta^2 = 0.03$), indicating that feedback valence significantly influenced the cross-stage pattern of change in behavioral trust. However, the three-way interaction among interaction stage, feedback source, and feedback valence did not reach significance ($F(4, 432) = 0.78, p = 0.539$).

The analysis of willingness for future cooperation showed that the main effect of interaction stage was significant ($F(4, 432) = 31.71, p < 0.001, partial\ \eta^2 = 0.23$), with willingness for future cooperation decreasing significantly as the interaction progressed; the main effect of feedback source was not significant ($F(1, 108) = 2.61, p = 0.109$); and the main effect of feedback valence was significant ($F(1, 108) = 38.18, p < 0.001, partial\ \eta^2 = 0.26$), with the level of willingness for future cooperation under the negative-feedback condition ($M = 4.10, SD = 1.82$) significantly lower than that under the positive-feedback condition ($M = 5.52, SD = 1.04$). The interaction between interaction stage and feedback source was significant ($F(4, 432) = 2.97, p = 0.019, partial\ \eta^2 = 0.03$), indicating that feedback source had a significant impact on the dynamic evolution of willingness for future cooperation; the interaction between interaction stage and feedback valence was significant ($F(4, 432) = 40.77, p < 0.001, partial\ \eta^2 = 0.27$), indicating that feedback valence had a significant impact on the dynamic evolution of willingness for future cooperation. The three-way interaction among interaction stage, feedback source, and feedback valence was marginally significant ($F(4, 432) = 2.31, p = 0.057, partial\ \eta^2 = 0.02$), suggesting that the pattern by which feedback source influenced the dynamic evolution of willingness for future cooperation may have differed across positive- and negative-feedback contexts.

4.4 Experimental Summary

Experiment 2 did not find that multiple rounds of positive feedback from robot versus human sources produced significant dynamic effects on behavioral trust or willingness for future cooperation; both indicators remained generally stable throughout the interaction process.

Combining the two experiments, under the paradigm and sample conditions of the present study, a clearer trend of differences in feedback source emerged in the negative-feedback context. This suggests that feedback valence may play a moderating role in the dynamic evolution of trust and cooperation, and that positive and negative feedback may produce asymmetric effects. It should be noted that, in the integrated analysis, the three-way interaction effect (interaction stage × feedback source × feedback valence) was not significant for behavioral trust and was only marginally significant for willingness for future cooperation. This may be partly due to the limited statistical power of the current sample size

for detecting complex interaction effects. Therefore, whether stable differences exist in the influence patterns of feedback source under positive- and negative-feedback contexts still requires further evidence from subsequent research.

5 Discussion

This study systematically examined the dynamic effects of multiple rounds of feedback from robot and human sources on trust and cooperation, revealing a distinctive pattern of trust evolution in human–robot collaboration and how it differs from interpersonal interaction. The study found that feedback from robot and human sources exerted differentiated effects on the dynamic evolution of trust and cooperation, and that this effect differed markedly across feedback-valence conditions. Specifically, in the context of multiple rounds of negative feedback, when the feedback source was a robot teammate, individuals' trust and cooperation exhibited a "lagged cumulative" pattern of evolution, characterized by delayed responses and cumulative effects; whereas when the feedback source was a human teammate, individuals' trust and cooperation exhibited an "immediate-response" pattern of evolution, characterized by high sensitivity and strong adaptability. However, in the context of multiple rounds of positive feedback, feedback from both robot and human sources did not significantly affect trust or cooperation.

5.1 Differentiated Evolutionary Patterns of Trust Under Negative Feedback: Lagged Accumulation and Immediate Response

The core finding of this study is that, under the influence of multiple rounds of negative feedback from robot and human sources, trust and cooperation showed

exhibit differentiated dynamic evolutionary patterns. Under the influence of multiple rounds of negative feedback from a robot source, individuals' trust and cooperation show a "lagged cumulative" evolutionary pattern: after the initial feedback, trust shows no significant change or only a small change, but as multiple rounds of feedback accumulate, a significant overall loss gradually emerges. By contrast, under conditions of multiple rounds of negative feedback from a human source, individuals' trust and cooperation show an "immediate response"evolutionary pattern: trust declines significantly immediately after the first round of negative feedback, but the subsequent downward trend becomes more gradual or even no longer shows significant change.

This finding is consistent with some views in human–robot interaction research, while also offering new and deeper insight. Paetzel et al. (2020) asked participants to play a geography game together with a robot and measured changes in participants' perceptions of the robot across three interactions, with three to ten days and no contact between each interaction. Based on their findings, they proposed that people' s perception of robot characteristics requires sufficient interaction time, and that a single interaction may not immediately change existing impressions. Rosén et al. (2024) designed a study in which par-

ticipants engaged in face-to-face conversational interaction with a robot, and during the process they measured participants' views of the robot' s affective and competence-related characteristics three times. Their results showed that people' s perceptions of robots possess a certain stability, and that newly experienced human–robot interaction is unlikely to immediately change inherent impressions and expectations of robots. These studies provide an important reference for understanding the relative lag in the influence of robot negative feedback in the present study.

Attribution Theory provides a possible theoretical explanatory pathway for the differentiated dynamic patterns observed in this study. The difference between the "lagged cumulative"and "immediate response"patterns observed here may be explained by individuals' different understandings of the intentions of feedback sources. Existing research shows that individuals' responses to evaluative feedback are, to a large extent, related to their attributions regarding the intentions of the feedback source, and that this attribution process may affect subsequent trust judgments (Desmet et al., 2011; Dirks et al., 2011). In human–robot interaction contexts, people may interpret robot feedback and human feedback differently (Guan Jian et al., 2025). Because robots are generally considered to lack sufficient autonomy and intentionality (Bigman & Gray, 2018; Gray et al., 2007), when individuals face negative feedback from a robot source, they may be more inclined to understand it as an expression of program settings, system rules, or objective standards (Yalcin et al., 2022), rather than as an interpersonal evaluation with a clear critical intention. Therefore, in the early stage of interaction, the impact of robot negative feedback on trust and cooperation may be relatively limited; however, as negative feedback is repeated, the negative information conveyed by the feedback content itself gradually accumulates, and ultimately may still lead to a significant decline in trust and cooperation. By comparison, in interpersonal interaction, others' words and behaviors are more readily understood as external expressions of their inner thoughts, emotions, or attitudes (Raver et al., 2012). Thus, when negative feedback comes from a human teammate, participants may be more likely to interpret it as subjective evaluation, dissatisfaction, or criticism, thereby showing a stronger decline in trust at an early stage.

In addition, the emotional responses elicited by different feedback sources may also constitute another theoretical pathway for understanding the above differences. Robot feedback may be less endowed with emotionality and interpersonal directedness, and therefore its immediate sense of threat is relatively low; human feedback, however, may more readily elicit emotional experiences such as shame, anger, hurt, or being negated, thereby affecting trust and cooperation. However, because this study did not directly measure attribution, perceived intentionality, or emotional responses, the above explanations remain possible explanatory pathways proposed on the basis of existing theories and research findings, and cannot be used to make definite causal inferences.

5.2 Potential asymmetry in the effects of positive and negative feedback

A comparison of the results of Experiment 1 and Experiment 2 suggests that the effects of positive and negative feedback on trust and cooperation may be asymmetric. In this study, the negative-feedback condition showed a clearer trend of source differences, whereas under the positive-feedback condition no apparent

···significant source differences. This result echoes Xuan and He' s (2025) finding regarding the moderating role of feedback valence, namely, that differentiated effects of feedback source may be more likely to emerge in negative-feedback rather than positive-feedback contexts. At the same time, this trend is also broadly consistent with the negative-bias perspective in trust dynamics (Yang et al., 2023).

Negative bias refers to the tendency for individuals to allocate more attention and cognitive resources to negative information; negative information also often has a stronger influence on judgment and behavior than positive information of comparable intensity (Baumeister et al., 2001; Carretié et al., 2001; Ito et al., 1998; Rozin & Royzman, 2001), and negative information is more likely to trigger changes in trust (Abele, 1985; Baumeister et al., 2001; Skowronski & Carlston, 1989; Srull & Wyer, 1989). Negative feedback may threaten individuals' self-esteem and sense of competence, prompting them to pay greater attention to whether the feedback provider has the intention to criticize or disparage them. Therefore, in negative-feedback contexts, differences between robots and humans in autonomy, intentionality, and interpersonal directedness may be more readily perceived, thereby further influencing trust updating. Positive feedback, by contrast, has a self-enhancing function; individuals may be more inclined to accept the content of the feedback directly and to engage less deeply in judging the intentions of the feedback source, such that source differences are relatively limited.

However, in the integrative cross-experiment analysis, the three-way interaction among interaction phase, feedback source, and feedback valence did not reach significance for behavioral trust, and reached only marginal significance for future willingness to cooperate. Thus, the present results are not yet sufficient to regard the asymmetry between positive and negative feedback as a stable effect that has been fully verified. A more prudent conclusion is that, under the paradigm and sample conditions of the present study, negative-feedback contexts exhibited more pronounced changes in trust and cooperation and a clearer trend of source differences; whether this trend can be generalized to other tasks, samples, and robot types remains to be further tested.

5.3 Theoretical Contributions and Practical Implications

This study has the following major theoretical contributions and directional practical implications:

First, this study extends research on human–machine trust from the dynamic perspective of sustained interaction. Previous research has mainly focused on the static "human–machine–environment" characteristic framework (Hancock et al., 2011; Khavas et al., 2020; Ulfert et al., 2024), involving three categories of factors: human-related characteristic factors, such as individual ability (Costa, 2003) and experience with technology use (Walter et al., 2015); robot-related characteristic factors, such as reliability (Hutchinson et al., 2023) and anthropomorphism (Kaplan et al., 2023); and environmental factors, such as task characteristics (Yang et al., 2023) and organizational culture (Fulmer & Gelfand, 2012). Although these studies have provided an important foundation for understanding human–machine trust, insufficient attention has been paid to the mechanisms through which interaction and communication operate in human–machine collaboration and to the dynamic evolution of trust. This study treats evaluative feedback as a processual factor that triggers trust updating and examines the dynamic changes in behavioral trust and future willingness to cooperate through five consecutive rounds of measurement, demonstrating that human–machine trust is manifested not only as differences in level at a given point in time, but may also be manifested as differences in rates of change and temporal trajectories.

Second, this study further reveals the temporal characteristics of how robot feedback and human feedback influence trust and cooperation. Previous research on human–machine feedback has mainly focused on the unidirectional influence of robot feedback on users, including its effects on employees' emotional states and collaborative experiences (Chen et al., 2024; Huang & Rau, 2024; You et al., 2011), as well as its effects on users' perceived competence, friendliness, and trustworthiness (Bracken et al., 2004; Mizumaru et al., 2019; Van der Hoorn et al., 2021). However, few studies have directly compared the differences between the effects of robot feedback and human feedback. We argue that, before robots are fully integrated into organizational environments, it is crucial to gain an in-depth understanding of the comparative advantages and limitations of human–machine collaborative teams relative to traditional human teams. Within a unified framework, this study systematically compared the dynamic effects of robot feedback and human feedback on trust and cooperation,

provided supplementary evidence for comparative research on human–robot collaboration and traditional human teams, and further distinguished the "lagged cumulative" evolutionary pattern of human–robot trust from the "immediate response" evolutionary pattern of interpersonal trust. This helps avoid simply regarding human–robot trust as an extension of interpersonal trust.

Finally, this study also offers directional practical implications for feedback design and trust maintenance in human–robot collaboration contexts. Based on the findings of this study, when future intelligent organizations introduce robots into feedback communication, they should attend to the differentiated psychological effects that may arise from different feedback sources. Specifically, robots may have potential value in organizational feedback systems for assisting in the

presentation of objective information and supporting standardized communication; however, whether they can buffer the immediate emotional responses caused by negative feedback still requires further examination. At the same time, human managers should continue to play an important role in contextual judgment, emotional support, meaning interpretation, and ethical responsibility during the feedback process. In addition, given that negative feedback from robots may exert a "lagged cumulative" influence, relevant practice should emphasize the dynamics of trust change during continuous interaction, rather than judging feedback effectiveness solely on the basis of immediate responses after a single interaction.

5.4 Research Limitations and Future Directions

Although this study has yielded rich findings, it still has the following limitations.

First, the participant sample was relatively homogeneous. The participants in this study were mainly university students, making it difficult to comprehensively reflect the response patterns of groups with different ages, educational levels, and work experience to robot feedback. In particular, considering Czaja et al.' s (2006) findings on the influence of age on technology acceptance, future research should include more diverse participant groups covering different age ranges, occupational backgrounds, and educational levels, so as to enhance the robustness and generalizability of the findings. In addition, although this study excluded individuals with a robotics-related educational background or prior experience using the Pepper robot, there were still differences in participants' familiarity with AI and robotics technologies, which may likewise have affected the results (Kopp et al., 2023). Therefore, future research could further examine the moderating roles of individual factors such as technological familiarity, general trust propensity toward automation, and personality dimensions in the effects of robot feedback.

Second, the ecological validity of the experimental task and research context has certain limitations. In this study, we designed a relatively comprehensive investment decision-making task based on real key indicators used in fund investment; however, many contextual factors that may influence feedback effects were still not fully reflected. Specifically: (1) the real consequences and level of risk were relatively low. Although team rankings and bonus rewards were set up, these differ substantially from high-stakes consequences in real organizations, where performance feedback is directly linked to pay, promotion, and the like; (2) task interdependence was relatively low. Although participants completed investment decisions together with teammates, this differs from collaborative contexts in real teams where members must coordinate closely and depend heavily on one another. At the same time, laboratory simulations cannot fully reflect the complexity of actual organizational environments. Future research should adopt methods with greater ecological validity, such as field studies and quasi-experimental designs, to explore in depth the influence of robot feedback

on the dynamic evolution of human–robot trust in real organizational settings and human–robot collaboration teams.

In addition, insufficient consideration was given to the influence of robots' own characteristics. Although the Pepper robot used in this study is currently a relatively advanced social robot, it may differ substantially from workplace robots in terms of appearance, interaction capabilities, autonomy, and other aspects. As emphasized by Skubis and Wodarski (2023), robotics technology is developing rapidly, and future robots may have stronger perception, decision-making, and execution capabilities, which will fundamentally change the basic patterns of human–robot interaction. Previous research has also pointed out that multiple robot characteristics, such as degree of anthropomorphism (Kulms & Kopp, 2016; Yam et al., 2022), voice and intonation (Damen & Toh, 2019; McDonnell & Baxter, 2019), and the expressive design of feedback (Groom et al., 2010), may have an effect on

people' s attitudes and responses. Therefore, future research should further explore how people perceive feedback from robots with different characteristics, as well as the specific effects of these characteristics on the dynamic evolution of human–robot trust.

Finally, this study primarily revealed the patterns by which multiple rounds of feedback from robotic and human sources influence the dynamic evolution of trust and cooperation. However, it did not directly test the psychological mechanisms underlying these patterns, nor did it measure potential mechanistic variables such as attribution style, perceived intentionality, and emotional responses. This limitation arose from trade-offs in the research design: (1) as a study examining the relationship between multiple rounds of feedback and dynamic trust in a human–robot context, we chose to prioritize description and testing at the phenomenological level, so as to provide an empirical foundation for subsequent mechanistic research; (2) this study required multi-time-point feedback manipulations and repeated measurements of dynamic trust, imposing a relatively heavy cognitive load on participants. If repeated measurements of mechanistic variables had been added, they might have impaired the fluency of the experimental process, and the measurement content itself might also have introduced interference effects. Therefore, future research should design experimental paradigms suitable for the dynamic measurement of a greater number of variables, in order to systematically test the psychological mechanisms by which multiple rounds of feedback influence the dynamic evolution of trust and cooperation.

6 Conclusion

Through one pilot experiment and two formal experiments, this study examined the dynamic effects of multiple rounds of positive and negative feedback from robotic and human sources on trust and cooperation, yielding the following main findings:

(1) In the context of multiple rounds of negative feedback, this study observed a trend toward source-based differences in how feedback from robots versus humans shaped the dynamic evolution of trust and cooperation. Specifically, when the feedback source was a robot teammate, individuals' trust and cooperation were more likely to exhibit a "lagged-accumulative" pattern of evolution, characterized by delayed responses and cumulative effects; whereas when the feedback source was a human teammate, individuals' trust and cooperation were more likely to exhibit an "immediate-response" pattern of evolution, characterized by higher sensitivity and stronger adaptability.

(2) Under the present paradigm and sample conditions, this study observed that trust and cooperation showed a clearer trend of source-based differences in the negative-feedback context, whereas no significant source-based difference was observed in the positive-feedback context. This result suggests that the effects of positive and negative feedback on trust and cooperation may be asymmetric, but this trend still requires further examination in future research.

(3) This study highlights the importance of investigating human–robot trust from a dynamic perspective. Although no obvious difference was found in the overall amount of change following robot feedback versus human feedback after multiple rounds of interaction, the dynamic patterns of trust evolution elicited by feedback during the collaborative process may still differ depending on the feedback source.

References

Aazam, M., Zeadally, S., & Harras, K. A. (2018). Deploying fog computing in industrial internet of things and industry 4.0. *IEEE Transactions on Industrial Informatics, 14*(10), 4674–4682. https://doi.org/10.1109/TII.2018.2855198

Abele, A. (1985). Thinking about thinking: Causal, evaluative and finalistic cognitions about social situations. *European Journal of Social Psychology, 15*(3), 315–332. https://doi.org/10.1002/ejsp.2420150306

Alarcon, G. M., Lyons, J. B., Hamdan, I. A., & Jessup, S. A. (2024). Affective responses to trust violations in a human-autonomy teaming context: Humans versus robots. *International Journal of Social Robotics, 16*(1), 23–35. https://doi.org/10.1007/s12369-023-01017-w

Ali, S. A. M., Said, N. A., Yunus, N. M., Kader, S. F. A., Latif, D. S. A., & Munap, R. (2014). Hackman and Oldham' s job characteristics model to job satisfaction. *Procedia - Social and Behavioral Sciences, 129*, 46–52. https://doi.org/10.1016/j.sbspro.2014.03.646

Aliasghari, P., Ghafurian, M., Nehaniv, C. L., & Dautenhahn, K. (2021). Effect of domestic trainee robots' errors on human teachers' trust. In *2021 30th*

IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp.81–88). Vancouver, Canada. https://doi.org/10.1109/RO-MAN50785.2021.9515510

Anseel, F., & Lievens, F. (2006). Certainty as a moderator of feedback reactions? A test of the strength of the self-verification motive. *Journal of Occupational and Organizational Psychology, 79*(4), 533–551. https://doi.org/10.1348/096317905X71462

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bracken, C. C., Jeffres, L. W., & Neuendorf, K. A. (2004). Criticism or praise? The impact of verbal versus text-only computer feedback on social presence, intrinsic motivation, and recall. *CyberPsychology & Behavior, 7*(3), 349–357. https://doi.org/10/c63wgk

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Candon, K., Zhou, H., Gillet, S., & Vázquez, M. (2023). Verbally soliciting human feedback in continuous human-robot collaboration: Effects of the framing and timing of reminders. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp.290–300). Stockholm, Sweden. https://doi.org/10.1145/3568162.3576980

Carretié, L., Mercado, F., Tapia, M., & Hinojosa, J. A. (2001). Emotion, attention, and the 'negativity bias' , studied through event-related potentials. *International Journal of Psychophysiology, 41*(1), 75–85. https://doi.org/10.1016/S0167-8760(00)00195-1

Chen, N., Cao, J., & Hu, X. (2024). The effects of robot managers' reward-punishment behaviors on human-robot trust and job performance. *International Journal of Social Robotics, 16*(3), 529–545. https://doi.org/10.1007/s12369-023-01091-0

Chi, V. B., & Malle, B. F. (2023). People dynamically update trust when interactively teaching robots. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp.554–564). Stockholm, Sweden. https://doi.org/10.1145/3568162.3576962

Cho, J. H., Chan, K., & Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR), 48*(2), 1–40. https://doi.org/10.1145/2815595

Christoffersen, K., & Woods, D. (2002). How to make automated systems team players. *Advances in Human Performance and Cognitive Engineering Research,*

2, 1–12. https://doi.org/10.1016/S1479-3601(02)02003-9

Costa, A. C. (2003). Work team trust and effectiveness. *Personnel Review, 32*(5), 605–622. https://doi.org/10.1108/00483480310488360

Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and Aging, 21*(2), 333–352. https://doi.org/10.1037/0882-7974.21.2.333

Damen, N., & Toh, C. (2019). Designing for trust: Understanding the role of agent gender and location on user perceptions of trust in home automation. *Journal of Mechanical Design, 141*(6), 061101. https://doi.org/10.1115/1.4042223

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour, 7*(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(3), 331–349. https://doi.org/10.1037/xap0000092

De Visser, E. J., Pak, R., & Neerincx, M. A. (2017). Trust development and repair in human-robot teams. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp.103–104). Vienna, Austria. https://doi.org/10.1145/3029798.3038409

De Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics, 12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Deepa, R., Sekar, S., Malik, A., Kumar, J., & Attri, R. (2024). Impact of AI-focussed technologies on social and technical competencies for HR managers – a systematic review and research agenda. *Technological Forecasting and Social Change, 202*, 123301. https://doi.org/10.1016/j.techfore.2024.123301

DeRosa, D. M., Hantula, D. A., Kock, N., & D' Arcy, J. (2004). Trust and leadership in virtual teamwork: A media naturalness perspective. *Human Resource Management, 43*(2–3), 219–232. https://doi.org/10.1002/hrm.20016

Desmet, P. T. M., De Cremer, D., & Van Dijk, E. (2011). Trust recovery following voluntary or forced financial compensations in the trust game: The role of trait forgiveness. *Personality and Individual Differences, 51*(3), 267–273. https://doi.org/10.1016/j.paid.2010.05.027

Dirks, K. T., Kim, P. H., Ferrin, D. L., & Cooper, C. D. (2011). Understanding the effects of substantive responses on trust following a transgression. *Organizational Behavior and Human Decision Processes, 114*(2), 87–103. https://doi.org/10.1016/j.obhdp.2010.10.003

Djamasbi, S., & Loiacono, E. T. (2008). Do men and women use feedback provided by their decision support systems (DSS) differently? *Decision Support Systems, 44*(4), 854–869.
https://doi.org/10.1016/j.dss.2007.10.008

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Fousiani, K. (2020). Power asymmetry, negotiations and conflict management in organizations. *Organizational conflict–New insights.* IntechOpen. https://doi.org/10.5772/intechopen.95492

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management, 38*(4), 1167–1230. https://doi.org/10.2139/ssrn.1873149

Getha-Taylor, H., Grayer, M. J., Kempf, R. J., & O' Leary, R. (2019). Collaborating in the absence of trust? What collaborative governance theory and practice can learn from the literatures of conflict resolution, psychology, and law. *The American Review of Public Administration, 49*(1), 51–64.

https://doi.org/10.1177/0275074018773089

Gkinko, L., & Elbanna, A. (2023). Designing trust: The formation of employees' trust in conversational AI in the digital workplace. *Journal of Business Research, 158*, 113707. https://doi.org/10.1016/j.jbusres.2023.113707

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. https://doi.org/10.1126/science.1134475

Groom, V., Chen, J., Johnson, T., Kara, F. A., & Nass, C. (2010). Critic, compatriot, or chump?: Responses to robot blame attribution. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 211–217). Osaka, Japan. https://doi.org/10.1109/HRI.2010.5453192

Gross, J., Leib, M., Offerman, T., & Shalvi, S. (2018). Ethical free riding: When honest people find dishonest partners. *Psychological Science, 29*(12), 1956–1968. https://doi.org/10.1177/0956797618796480

Grossman, R., & Feitosa, J. (2018). Team trust over time: Modeling reciprocal and contextual influences in action teams. *Human Resource Management Review, 28*(4), 395–410. https://doi.org/10.1016/j.hrmr.2017.03.006

Guan, J., Li, W. P., He, G. H., & Zhang, X. A. (2025). Artificial intelligence feedback: Literature review and research prospects. *Foreign Economics & Management, 47*(3), 83–100. https://doi.org/10.16538/j.cnki.fem.20240622.301

[Guan, J., Li, W. P., He, G. H., & Zhang, X. A. (2025). Artificial intelligence feedback: Literature review and research prospects. *Foreign Economics & Management, 47*(3), 83–100. https://doi.org/10.16538/j.cnki.fem.20240622.301]

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial intelligence trust, risk and security management (AI trism): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications, 240*, 122442. https://doi.org/10.1016/j.eswa.2023.122442

Haenlein, M., & Kaplan, A. (2021). Artificial intelligence and robotics: Shaking up the business world and society at large. *Journal of Business Research, 124*, 405–407. https://doi.org/10.1016/j.jbusres.2020.10.042

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hecklau, F., Galeitzke, M., Flachs, S., & Kohl, H. (2016). Holistic approach for human resource management in Industry 4.0. *Procedia CIRP, 54*, 1–6. https://doi.org/10.1016/j.procir.2016.05.102

He, G. B., Chen, C., He, Z. T., Cui, L. D., Lu, J. Q., Xuan, H. Z., & Lin, L. (2022). Human-agent collaborative decision-making in intelligent organizations: A perspective of human-agent inner compatibility. *Advances in Psychological Science, 30*(12), 2619–2627.

[He, G. B., Chen, C., He, Z. T., Cui, L. D., Lu, J. Q., Xuan, H. Z., & Lin, L. (2022). Human-agent collaborative decision-making in intelligent organizations: A study based on human-agent internal compatibility. *Advances in Psychological Science, 30*(12), 2619–2627. https://doi.org/10.3724/SP.J.1042.2022.02619]

Hicks, L. J., Smith, A. C., Ralph, B. C. W., & Smilek, D. (2020). Restoration of sustained attention following virtual nature exposure: Undeniable or unreliable? *Journal of Environmental Psychology, 71*, 101488. https://doi.org/10.1016/j.jenvp.2020.101488

Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the message across: The problem of communicating assessment feedback. *Teaching in Higher Education, 6*(2), 269–274. https://doi.org/10.1080/13562510120045230

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.

https://doi.org/10.1177/0018720814547570

Huang, H., & Rau, P. L. P. (2024). How to provide feedback? The role of robot' s language and feedback framework. *International Journal of Human-Computer Interaction, 40*(8), 1856–1872.

https://doi.org/10.1080/10447318.2023.2223946

Hutchinson, J., Strickland, L., Farrell, S., & Loft, S. (2023). The perception of automation reliability and acceptance of automated advice. *Human Factors, 65*(8), 1596–1612.
https://doi.org/10.1177/00187208211062985

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology, 75*(4), 887–900. https://doi.org/10.1037/0022-3514.75.4.887

Jawahar, I. M. (2006). Correlates of satisfaction with performance appraisal feedback. *Journal of Labor Research, 27*(2), 213–236. https://doi.org/10.1007/s12122-006-1004-1

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review, 23*(3), 531–546. https://doi.org/10.5465/amr.1998.926625

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Khavas, Z. R., Ahmadzadeh, S. R., & Robinette, P. (2020). Modeling trust in human-robot interaction: A survey. *International Conference on Social Robotics*, 529–541. Springer International Publishing. https://doi.org/10.1007/978-3-030-62056-1_44

Kinicki, A. J., Prussia, G. E., Wu, B. J., & McKee-Ryan, F. M. (2004). A covariance structure analysis of employees' response to performance feedback. *Journal of Applied Psychology, 89*(6), 1057–1069. https://doi.org/10.1037/0021-9010.89.6.1057

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly, 30*(4), 941–960. https://doi.org/10.2307/25148760

Kopp, T., Baumgartner, M., & Kinkel, S. (2023). "It' s not Paul, it' s a robot" : The impact of linguistic framing and the evolution of trust and distrust in a collaborative robot during a human-robot interaction. *International Journal of Human-Computer Studies, 178*, 103095. https://doi.org/10.1016/j.ijhcs.2023.103095

Korsgaard, M. A., Kautz, J., Bliese, P., Samson, K., & Kostyszyn, P. (2018). Conceptualising time as a level of analysis: New directions in

the analysis of trust dynamics. *Journal of Trust Research, 8*(2), 142–165. https://doi.org/10.1080/21515581.2018.1516557

Kożuch, B., & Sienkiewicz-Małyjurek, K. (2022). Building collaborative trust in public safety networks. *Safety Science, 152*, 105785. https://doi.org/10.1016/j.ssci.2022.105785

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors, 62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Kuhnen, C. M., & Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science, 58*(1), 94–113. https://doi.org/10.1287/mnsc.1110.1379

Kulms, P., & Kopp, S. (2016). The effect of embodiment and competence on trust and cooperation in human-agent interaction. In *Intelligent Virtual Agents: 16th International Conference* (pp.75–84). Los Angeles, United States. Springer International Publishing. https://doi.org/10.1007/978-3-319-47665-0_7

Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. In *Proceedings of Mensch und Computer 2019* (pp.31–42). Hamburg, Germany. https://doi.org/10.1145/3340764.3340793

Kuvaas, B. (2006). Performance appraisal satisfaction and employee outcomes: Mediating and moderating roles of work motivation. *The International Journal of Human Resource Management, 17*(3), 504–522. https://doi.org/10.1080/09585190500521581

Lanz, L., Briker, R., & Gerpott, F. H. (2024). Employees adhere more to unethical instructions from human than AI supervisors: Complementing experimental evidence with machine learning. *Journal of Business Ethics,*

189(3), 625–646. https://doi.org/10.1007/s10551-023-05393-1

Lee, D., Stajkovic, A. D., & Cho, B. (2011). Interpersonal trust and emotion as antecedents of cooperation: Evidence from korea. *Journal of Applied Social Psychology, 41*(7), 1603–1631. https://doi.org/10.1111/j.1559-1816.2011.00776.x

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 2053951718756684. https://doi.org/10.1177/2053951718756684

Leung, K., Su, S., & Morris, M. W. (2001). When is criticism not constructive? The roles of fairness perceptions and dispositional attributions in employee ac-

ceptance of critical supervisory feedback. *Human Relations, 54*(9), 1155–1187. https://doi.org/10.1177/0018726701549002

Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied, 6*(2), 104–123. https://doi.org/10.1037/1076-898X.6.2.104

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces, 63*(4), 967–985. https://doi.org/10.1093/sf/63.4.967

Li, J.-M., Zhang, R.-X., Wu, T.-J., & Mao, M. (2024). How does work autonomy in human-robot collaboration affect hotel employees' work and health outcomes? Role of job insecurity and person-job fit. *International Journal of Hospitality Management, 117*, 103654. https://doi.org/10.1016/j.ijhm.2023.103654

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Luo, R., Du, N., & Yang, X. J. (2022). Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human-Computer Interaction, 38*(18–20), 1962–1971. https://doi.org/10.1080/10447318.2022.2097602

Lv, X., Shi, K., He, Y., Ji, Y., & Lan, T. (2024). My colleague is not "human" : Will working with robots make you act more indifferently? *Journal of Business Research, 176*, 114585. https://doi.org/10.1016/j.jbusres.2024.114585

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10/d4sv4f

Malik, A., Budhwar, P., & Kazmi, B. A. (2023). Artificial intelligence (AI)-assisted HRM: Towards an extended strategic framework. *Human Resource Management Review, 33*(1), 100940. https://doi.org/10.1016/j.hrmr.2022.100940

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review, 26*(3), 356–376. https://doi.org/10.5465/amr.2001.4845785

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Journal of The Academy of Management, 38*(1), 24–59.

McDonnell, M., & Baxter, D. (2019). Chatbots and gender stereotyping. *Interacting with Computers, 31*(2), 116–121. https://doi.org/10.1093/iwc/iwz007

Mizumaru, K., Satake, S., Kanda, T., & Ono, T. (2019). Stop doing it! Approaching strategy for a robot to admonish pedestrians. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp.449–457). Daegu, Korea (South). https://doi.org/10.1109/HRI.2019.8673017

Mou, Y., & Xu, K. (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior, 72*, 432–440. https://doi.org/10/gf2k3r

Narzary, S., Makhecha, U. P., Budhwar, P., Malik, A., & Kumar, S. (2024). A temporal evolution of human resource management and technology research: A retrospective bibliometric analysis. *Personnel Review, 54*(3), 800–823. https://doi.org/10.1108/PR-04-2023-0296

Nicholson, C. Y., Compeau, L. D., & Sethi, R. (2001). The role of interpersonal liking in building trust in long-term channel relationships. *Journal of the Academy of Marketing Science, 29*(1), 3–15. https://doi.org/10.1177/0092070301291001

O' Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human-autonomy teaming: A review and analysis of the empirical literature. *Human Factors, 64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013). Building appropriate trust in human-robot teams. In *AAAI Spring Symposium: Trust and Autonomous Systems* (pp. 60–65), California, United States.

Ozaras, G., & Abaan, S. (2018). Investigation of the trust status of the nurse-patient relationship. *Nursing Ethics, 25*(5), 628–639. https://doi.org/10.1177/0969733016664971

Paetzel, M., Perugia, G., & Castellano, G. (2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp.73–82). Cambridge, United Kingdom. https://doi.org/10.1145/3319502.3374786

Pereira, V., Hadjielias, E., Christofi, M., & Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective. *Human Resource Management Review, 33*(1), 100857. https://doi.org/10.1016/j.hrmr.2021.100857

Podsakoff, P. M., Bommer, W. H., Podsakoff, N. P., & MacKenzie, S. B. (2006). Relationships between leader reward and punishment behavior and subordinate attitudes, perceptions, and behaviors: A meta-analytic review of existing and new research. *Organizational Behavior and Human Decision Processes, 99*(2), 113–142. https://doi.org/10.1016/j.obhdp.2005.09.002

Qi, Y., Chen, J. T., Qin, S. T., & Du, F. (2024). Human-AI mutual trust in the

era of artificial general intelligence. *Advances in Psychological Science, 32*(12), 2124-2136. https://doi.org/10.3724/SP.J.1042.2024.001

[Qi Yue, Chen Junting, Qin Shaotian, Du Feng. (2024). Human-AI trust in the era of artificial general intelligence. *Advances in Psychological Science, 32*(12), 2124-2136. https://doi.org/10.3724/SP.J.1042.2024.001]

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Raver, J. L., Jensen, J. M., Lee, J., & O' Reilly, J. (2012). Destructive criticism revisited: Appraisals, task outcomes, and the moderating role of competitiveness. *Applied Psychology, 61*(2), 177–203. https://doi.org/10.1111/j.1464-0597.2011.00462.x

Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems, 47*(4), 425–436. https://doi.org/10.1109/THMS.2017.2648849

Roesler, E., Vollmann, M., Manzey, D., & Onnasch, L. (2024). The dynamics of human-robot trust attitude and behavior—exploring the effects of anthropomorphism and type of failure. *Computers in Human Behavior, 150*, 108008. https://doi.org/10.1016/j.chb.2023.108008

Rosén, J., Lindblom, J., Lamb, M., & Billing, E. (2024). Previous experience matters: An in-person investigation of expectations in human-robot interaction. *International Journal of Social Robotics, 16*(3), 447–460.

https://doi.org/10.1007/s12369-024-01107-3

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Human Factors*, *61*(4), 614–626. https://doi.org/10.1177/0018720818816838

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Skubis, I., & Wodarski, K. (2023). Humanoid robots in managerial positions –decision-making process and human oversight. *Scientific Papers of Silesian University of Technology. Organization and Management Series*, *2023*(189), 573–596. https://doi.org/10.29119/1641-3466.2023.189.36

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review*, *96*(1), 58–83. https://doi.org/10.1037/0033-295X.96.1.58

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320–332. https://doi.org/10.1037/0021-9010.87.2.320

Ulfert, A.-S., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2024). Shaping a multidisciplinary understanding of team trust in human-AI teams: A theoretical framework. *European Journal of Work and Organizational Psychology*, *33*(2), 158–171. https://doi.org/10.1080/1359432X.2023.2200172

Van der Hoorn, D. P. M., Neerincx, A., & de Graaf, M. M. A. (2021). "I think you are doing a bad job!" : The effect of blame attribution by a robot in human-robot collaboration. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp.140–148). Boulder, United States. https://doi.org/10/gjgptk

Van Dongen, K., & Van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, *71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

Van Pinxteren, M. M. E., Wetzels, R. W. H., Rüger, J., Pluymaekers, M., & Wetzels, M. (2019). Trust in humanoid robots: Implications for services marketing. *Journal of Services Marketing*, *33*(4), 507–518. https://doi.org/10.1108/JSM-01-2018-0045

Vance, A., Elie-Dit-Cosaque, C., & Straub, D. W. (2008). Examining trust in information technology artifacts: The effects of system quality and culture. *Journal of Management Information Systems*, *24*(4), 73–100. https://doi.org/10.2753/MIS0742-1222240403

Vanneste, B. S., & Puranam, P. (2024). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review*, *50*(4), 726–744. https://doi.org/10.5465/amr.2022.0041

Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, *4*(4), 404–409. https://doi.org/10.5465/amd.2018.0084

Walter, N., Ortbach, K., & Niehaves, B. (2015). Designing electronic feedback -analyzing the effects of social presence on perceived feedback usefulness. *International Journal of Human-Computer Studies*, *76*, 1–11. https://doi.org/10.1016/j.ijhcs.2014.12.001

Watts, S. A. (2007). Evaluative feedback: Perspectives on media effects. *Journal of Computer-Mediated Communication*, *12*(2), 384–411. https://doi.org/10.1111/j.1083-6101.2007.00330.x

Weiss, H. M., Suckow, K., & Cropanzano, R. (1999). Effects of justice conditions on discrete emotions. *Journal of Applied Psychology*, *84*(5), 786–794. https://doi.org/10/dp7s56

Whitehair, C., & Berdanier, C. G. (2017). The role of trust in collaborative research settings: Opportunities for future research in graduate engineering education. *2017 ASEE Annual Conference & Exposition*, Columbus, United States. https://doi.org/10.18260/1-2-29008

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science, 19*(3), 353–374. https://doi.org/10/ggxp4d

Xu, L. Y., Yu, F., & Peng, K. P. (2022). Algorithmic discrimination causes less desire for moral punishment than human discrimination. *Acta Psychologica Sinica, 54*(9), 1076–1092. https://doi.org/10.3724/SP.J.1041.2022.01076

[Xu Liying, Yu Feng, & Peng Kaiping. (2022). Algorithmic discrimination elicits less desire for moral punishment than human discrimination. *Acta Psychologica Sinica, 54*(9), 1076–1092. https://doi.org/10.3724/SP.J.1041.2022.01076]

Xu, M., David, J. M., & Kim, S. H. (2018). The fourth industrial revolution: Opportunities and challenges. *International Journal of Financial Research, 9*(2), 90–95. https://doi.org/10.5430/ijfr.v9n2p90

Xuan, H., & He, G. (2025). Negative feedback from robots is received better than that from humans: The effect of feedback on human-robot trust and collaboration. *Journal of Business Research, 193*, 115333. https://doi.org/10.1016/j.jbusres.2025.115333

Yalcin, G., Lim, S., Puntoni, S., & Van Osselaer, S. M. J. (2022). Thumbs up or down: Consumer reactions to decisions by algorithms versus humans. *Journal of Marketing Research, 59*(4), 696–717. https://doi.org/10.1177/00222437211070016

Yam, K. C., Goh, E.-Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology, 102*, 104360. https://doi.org/10.1016/j.jesp.2022.104360

Yang, X. J., Schemanske, C., & Searle, C. (2023). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors, 65*(5), 862–878. https://doi.org/10.1177/00187208211034716

You, S., Nie, J., Suh, K., & Sundar, S. S. (2011). When the robot criticizes you···Self-serving bias in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 295–296). Lausanne, Switzerland. https://doi.org/10/b7td8d

Dynamic Effects of Multiround Positive and Negative Feedback from Robots and Humans on Trust and Collaboration

XUAN Hongzhou1, 2, 3, ZHU Jiaying2, LAI Qianen2, HE Guibing2

(1 *School of Business Administration, Zhejiang Gongshang University, Hangzhou 310018, China*)

(2 *Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310058, China*)

(3 Innovation Center for Business Technology Application, *Zhejiang Gongshang University, Hangzhou 310018, China*)

Abstract

Human–robot collaboration is increasingly common in intelligent organizations, making it important to understand how trust changes across repeated interactions. Evaluative feedback can trigger trust updating, yet prior research has largely relied on single-time-point measures or one-off feedback. This research examined the dynamic effects of multiround positive and negative feedback from robot and human teammates on behavioral trust and willingness to collaborate.

A pilot study and two laboratory experiments were conducted. In the pilot study, 192 adults evaluated candidate feedback statements, from which four positive and four negative statements of comparable intensity were selected. In Experiment 1, 56 participants collaborated with either a Pepper robot or a human teammate and received four rounds of negative feedback. Experiment 2 used the same design with another 56 participants but presented positive feedback. In both experiments, participants completed five rounds of an investment decision task. Behavioral trust was indexed by the extent to which participants revised their judgments toward the teammate' s recommendation, whereas willingness to collaborate was measured after each round.

Under multiround negative feedback, behavioral trust and willingness to collaborate declined over time, but their trajectories differed by feedback source. Feedback from the robot teammate produced a relatively gradual and delayed pattern: behavioral trust changed little initially but declined in later rounds, while willingness to collaborate decreased as feedback accumulated. Feedback from the human teammate produced a sharper early decline followed by relative stabilization. Segmented models fit the data better than linear or quadratic models, supporting a lagged-cumulative pattern for robot feedback and an immediate-responsive pattern for human feedback.

These findings suggest that robots and humans may produce similar overall declines under negative feedback but different temporal processes of trust updating. Human feedback appears to have a stronger immediate interpersonal impact,

whereas robot feedback may influence trust through accumulation. The results highlight the importance of examining when and how trust changes rather than relying only on end-point comparisons. The apparent asymmetry between positive and negative feedback should be interpreted cautiously because the higher-order interaction involving feedback valence was not consistently significant.

Keywords: human-robot collaboration, behavioral trust, evaluative feedback, trust dynamics, feedback valence

Appendix A Parameterization, Piecewise Specification, and Supplementary Analysis of the Mixed-Effects Models

A1 General Structure of the Mixed-Effects Models

To characterize the dynamic changes in participants' behavioral trust and future willingness to cooperate during repeated interactions, this study used linear mixed-effects models. The general form of the model is as follows:

$$Y_{ij} = \beta_0 + \sum_k \beta_k X_{k,i,j} + u_{0i} + u_{1i}\text{Time}_{ij} + \epsilon_{ij}$$

where Y_{ij} denotes the outcome variable for participant i at time point j (behavioral trust or future willingness to cooperate); β_0 is the fixed intercept; $X_{k,i,j}$ denotes time, feedback source, their interaction, or piecewise time variables; u_{0i} and u_{1i} denote, respectively, the participant-level random intercept and random time slope, used to characterize individual differences in initial level and in trajectories of change over time; and the residual term ϵ_{ij} is assumed to follow a normal distribution with mean 0.

All models were estimated using maximum likelihood and included the random-effects structure $(1 + \text{Time} \mid \text{Subject})$.

A2 Parameterization of the Three Specific Models

A2.1 Linear Mixed-Effects Model

The linear model was used to test whether the outcome variable showed monotonic change across interaction phases. Its form was:

$$Y_{ij} = \beta_0 + \beta_1\text{Time}_{ij} + \beta_2\text{Source}_i + \beta_3(\text{Time}_{ij} \times \text{Source}_i) + u_{0i} + u_{1i}\text{Time}_{ij} + \epsilon_{ij}$$

where:

- Y_{ij} is the dependent variable for participant i at time point j (behavioral trust or future willingness to cooperate)
- $\text{Time}_{ij} \in \{0, 1, 2, 3, 4\}$ denotes the interaction phase
- Source_i is the effect coding of the feedback source (robot group = 0.5, human group = −0.5)
- β_0 is the intercept
- β_1 is the main effect of time
- β_2 is the main effect of source
- β_3 is the time × source interaction effect
- $u_{0i} \sim N(0, \tau_0^2)$ is the participant random intercept
- $u_{1i} \sim N(0, \tau_1^2)$ is the participant random slope
- $\epsilon_{ij} \sim N(0, \sigma^2)$ is the residual term
- $\text{Cov}(u_{0i}, u_{1i}) = \tau_{01}$

A2.2 Quadratic Mixed-Effects Model

To test nonlinear trends in change over time, a quadratic term for time was added to the model described above:

$$Y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Time}_{ij}^2 + \beta_3 \text{Source}_i + \beta_4 (\text{Time}_{ij} \times \text{Source}_i) + \beta_5 (\text{Time}_{ij}^2 \times \text{Source}_i) + u_{0i} + u_{1i} \text{Time}_{ij} + \epsilon_{ij}$$

where:

- β_2 is the coefficient of the quadratic term
- β_5 is the quadratic term × source interaction effect
- The remaining parameters are defined as in Model 1.

A2.3 Segmented Mixed-Effects Model

To further characterize the potentially different mechanisms of change in the early and later stages of interaction, this study constructed a segmented model. Breakpoint selection was based on the following considerations: (1) theoretically, the initial stage of interaction is regarded as a critical phase of rapid trust adjustment and expectation revaluation (Dirks et al., 2011); (2) the descriptive statistics and mixed ANOVA results indicate that there are certain differences in the change trends between T0–T1 and the subsequent stages. Therefore, this study used T1 as the breakpoint of the segmented model, dividing time into an early stage (T0–T1) and a later stage (T2–T4). The complete form of the segmented model is as follows:

$$Y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij}^{\text{R,}} + \beta_2 \text{Time}_{ij}^{\text{R,}} + \beta_3 \text{Time}_{ij}^{\text{H,}} + \beta_4 \text{Time}_{ij}^{\text{H,}} + u_{0i} + u_{1i} \text{Time}_{ij} + \epsilon_{ij}$$

where the segmented time variables are defined as follows:

Robot group:

$$\text{Time}_{ij}^{\text{R,}} = \begin{cases} \text{Time}_{ij}, & \text{if Time}_{ij} \le 1 \\ 1, & \text{if Time}_{ij} > 1 \end{cases}$$

$$\text{Time}_{ij}^{\text{R,}} = \begin{cases} 0, & \text{if Time}_{ij} \le 1 \\ \text{Time}_{ij} - 1, & \text{if Time}_{ij} > 1 \end{cases}$$

Human group:

$$\text{Time}_{ij}^{\text{H,}} = \begin{cases} \text{Time}_{ij}, & \text{if Time}_{ij} \le 1 \\ 1, & \text{if Time}_{ij} > 1 \end{cases}$$

$$\text{Time}_{ij}^{\text{H,}} = \begin{cases} 0, & \text{if Time}_{ij} \le 1 \\ \text{Time}_{ij} - 1, & \text{if Time}_{ij} > 1 \end{cases}$$

Parameter interpretation:

- β_1: early-stage slope for the robot group (T0→T1)
- β_2: later-stage slope for the robot group (T2→T4)
- β_3: early-stage slope for the human group (T0→T1)
- β_4: later-stage slope for the human group (T2→T4)

A2.4 Sensitivity Analysis of the Segmentation Breakpoint

To examine the influence of breakpoint setting on the segmented-model results for Experiment 1, we further compared the goodness of fit of segmented models with different breakpoint positions (T1–T3). The results show that, for both behavioral trust and future willingness to cooperate, the model corresponding to the current breakpoint (T1) is superior to the other breakpoint settings on both the AIC and BIC indices.

Table A1. Comparison of goodness of fit of segmented models under different breakpoint settings in Experiment 1 (AIC / BIC)

Breakpoint	Behavioral Trust AIC	Behavioral Trust BIC	Future Willingness to Cooperate AIC	Future Willingness to Cooperate BIC
T1	−181.33	−148.62	806.99	839.70
T2	−179.94	−147.23	808.95	841.67
T3	−174.19	−141.47	826.91	859.62

A3 Conditional Estimation and Visualization of Marginal Effects

Based on the model results above, we further calculated the marginal mean estimates over time under different feedback-source conditions, and plotted the conditional-effect curves, as shown in Figure A1 and Figure A2.

Figure A1. Marginal-effect estimates of behavioral trust over time under different feedback-source conditions

In-figure labels: y-axis, "Behavioral trust" ; x-axis, "Interaction stage" (`T0`, `T1`, `T2`, `T3`, `T4`); legend, "Feedback source" : solid line = "Human," dashed line = "Robot."

Note: The curves indicate the model-predicted marginal means, and the shaded regions indicate 95% confidence intervals.

Figure A2. Marginal-effect estimates of future willingness to cooperate over time under different feedback-source conditions

In-figure labels: y-axis, "Future willingness to cooperate" ; x-axis, "Interaction stage" (`T0`, `T1`, `T2`, `T3`, `T4`); legend, "Feedback source" : solid line = "Human," dashed line = "Robot."

Note: The curves indicate the model-predicted marginal means, and the shaded regions indicate 95% confidence intervals.

Appendix B Generalized η^2 and ω^2 for Each Effect in the Mixed-Methods Analysis

B1 Generalized η^2 and ω^2 for Each Effect in Experiment 1

B1.1 Behavioral Trust Indicator

- Feedback source:

$$F(1, 54) = 0.01, p = 0.947, \text{partial } \eta^2 < 0.001, 95\% \text{ CI } [0.00, 0.03],$$

$$\text{generalized } \eta^2 < 0.001, \omega^2 < 0.001$$

- Interaction phase:

$$F(4, 216) = 7.43, p < 0.001, \text{partial } \eta^2 = 0.12, 95\% \text{ CI } [0.04, 0.20],$$

$$\text{generalized } \eta^2 = 0.07, \omega^2 = 0.10$$

- Feedback source × interaction phase:

$$F(4, 216) = 2.30, p = 0.060, \text{partial } \eta^2 = 0.04, 95\% \text{ CI } [0.00, 0.09],$$

$$\text{generalized } \eta^2 = 0.02, \omega^2 = 0.02$$

B1.2 Future Willingness-to-Cooperate Indicator

- Feedback source:

$$F(1, 54) = 1.33, p = 0.254, \text{partial } \eta^2 = 0.02, 95\% \text{ CI } [0.00, 0.16],$$

$$\text{generalized } \eta^2 = 0.02, \omega^2 = 0.01$$

- Interaction phase:

$$F(4, 216) = 54.12, p < 0.001, \text{partial } \eta^2 = 0.50, 95\% \text{ CI } [0.39, 0.58],$$

$$\text{generalized } \eta^2 = 0.17, \omega^2 = 0.49$$

- Feedback source × interaction phase:

$$F(4, 216) = 3.70, p = 0.006, \text{partial } \eta^2 = 0.06, 95\% \text{ CI } [0.01, 0.12],$$

$$\text{generalized } \eta^2 = 0.01, \omega^2 = 0.05$$

B2 Generalized η^2 and ω^2 for Each Effect in Experiment 2

B2.1 Behavioral Trust Indicator

- Feedback source:

$$F(1, 54) = 1.70, p = 0.198, \text{partial } \eta^2 = 0.03, 95\% \text{ CI } [0.00, 0.17],$$

$$\text{generalized } \eta^2 = 0.01, \omega^2 = 0.01$$

- Interaction phase:

$$F(4, 216) = 0.50, p = 0.739, \text{partial } \eta^2 = 0.01, 95\% \text{ CI } [0.00, 0.03],$$

$$\text{generalized } \eta^2 = 0.01, \omega^2 < 0.001$$

- Feedback source × interaction phase:

$$F(4, 216) = 0.47, p = 0.761, \text{partial } \eta^2 = 0.01, 95\% \text{ CI } [0.00, 0.03],$$

$$\text{generalized } \eta^2 = 0.01, \omega^2 < 0.001$$

B2.2 Future Willingness-to-Cooperate Indicator

- Feedback source:

$$F(1, 54) = 1.48, p = 0.229, \text{partial } \eta^2 = 0.03, 95\% \text{ CI } [0.00, 0.16],$$

$$\text{generalized } \eta^2 = 0.02,\ \omega^2 = 0.01$$

- Interaction phase:

$$F(4, 216) = 1.45,\ p = 0.218,\ \text{partial } \eta^2 = 0.03,\ 95\% \text{ CI } [0.00, 0.07],$$

$$\text{generalized } \eta^2 = 0.01,\ \omega^2 = 0.01$$

- Feedback source × interaction phase:

$$F(4, 216) = 0.57,\ p = 0.686,\ \text{partial } \eta^2 = 0.01,\ 95\% \text{ CI } [0.00, 0.03],$$

$$\text{generalized } \eta^2 = 0.003,\ \omega^2 < 0.001$$

Appendix C Exploratory Analysis of Feedback Acceptance

C1 Definition and Measurement Method

C1.1 Definition

Feedback acceptance refers to an individual's perception of the accuracy, fairness, and effectiveness of feedback, reflecting the individual' s most direct evaluation of the feedback (Jawahar, 2006; Kinicki et al., 2004; Kuvaas, 2006; Leung et al., 2001).

C1.2 Measurement Method

After each instance of feedback was received, participants were immediately asked to evaluate their acceptance of the feedback on a 7-point Likert scale, ranging from "completely unacceptable" to "completely acceptable." In each experiment, the teammate provided feedback four times; therefore, each participant completed four measurements of feedback acceptance (after the first feedback, T1; after the second feedback, T2; after the third feedback, T3; and after the fourth feedback, T4).

C2 Results of the Exploratory Analysis

This experiment conducted statistical testing using a 2 (feedback source) × 4 (interaction stage) mixed analysis of variance (mixed ANOVA), with interaction stage treated as a repeated-measures factor. At the same time, the Holm method ($\alpha = 0.05$) was used to correct for multiple comparisons in the simple-effects analyses.

C2.1 Dynamic Evolution of Feedback Acceptance in Experiment 1

Figure labels: y-axis, "Feedback acceptance"; x-axis, "Interaction stage"; legend title, "Feedback source" ; legend entries, "Human" and "Robot" ; x-axis tick labels, T1, T2, T3, T4.

Figure C1. Dynamic evolution of feedback acceptance under multiple rounds of negative feedback

Note: Markers indicate means, and error bands indicate standard errors.

The dynamic evolution of feedback acceptance under continuous negative feedback is shown in Figure C1. The results of the mixed ANOVA indicated that the main effect of feedback source was not significant ($F(1, 54) = 0.652$, $p = 0.423$): overall feedback acceptance toward the robot teammate ($M = 4.34$, $SD = 1.45$) did not differ significantly from overall feedback acceptance toward the human teammate ($M = 4.04$, $SD = 1.47$). The main effect of interaction stage was significant ($F(3, 162) = 7.267$, $p < 0.001$, partial $\eta^2 = 0.119$), with feedback

acceptance decreasing significantly from the T1 stage ($M = 4.50$, $SD = 1.35$) to the T4 stage ($M = 3.98$, $SD = 1.49$). The interaction between feedback source and interaction stage was significant ($F(3, 162) = 5.665$, $p = 0.001$, partial $\eta^2 = 0.095$), suggesting that the pattern of change in feedback acceptance across stages may differ under different feedback-source conditions. Further simple-effects analyses showed that, under the robot feedback-source condition

under which feedback acceptance declined significantly from the T1 stage ($M = 4.91, SD = 0.82$) to the T2 stage ($M = 4.41, SD = 1.26$) ($p_{\text{Holm}} = 0.021$), with no significant changes between subsequent adjacent stages ($p_{\text{Holm}} > 0.785$); whereas under the human-feedback-source condition, there were no significant changes between adjacent stages ($p_{\text{Holm}} = 1.000$). Notably, at the T1 stage, participants' acceptance of feedback from robot teammates ($M = 4.91, SD = 0.82$) was significantly higher than that of feedback from human teammates ($M = 4.10, SD = 1.64; p_{\text{Holm}} = 0.023$); whereas at the subsequent T2, T3, and T4 stages, there were no significant differences in participants' acceptance of feedback from robot teammates versus human teammates ($p_{\text{Holm}}s > 0.323$).

C2.2 Dynamic Evolution of Feedback Acceptance in Experiment 2

[Line chart: y-axis, "Feedback acceptance" ; x-axis, "Interaction stage" with T1, T2, T3, and T4. Legend title: "Feedback source" ; series: "Human" and "Robot."]

Figure C2. Dynamic evolution of feedback acceptance under multiple rounds of positive feedback

Note: Markers indicate means, and error bands indicate standard errors.

The process by which feedback acceptance dynamically evolved with continuous positive feedback is shown in Figure C2. The mixed-design ANOVA results showed that the main effect of feedback source was not significant ($F(1, 54) = 0.205, p = 0.653$): overall feedback acceptance for robot teammates ($M = 5.38, SD = 1.00$) did not differ significantly from overall feedback acceptance for human teammates ($M = 5.27, SD = 1.01$). The main effect of interaction stage was not significant ($F(3, 162) = 2.287, p = 0.081$): feedback acceptance remained relatively stable overall from the T1 stage ($M = 5.18, SD = 0.87$) to the T4 stage ($M = 5.27, SD = 1.05$). The interaction between feedback source and interaction stage was likewise not significant ($F(3, 162) = 2.076, p = 0.105$).

C2.3 Summary

Under the multiple-round negative-feedback condition, participants' acceptance of feedback from robot teammates was significantly higher than that of feedback from human teammates at the first instance of negative feedback, but then gradually declined until it no longer differed from the human group. This pattern is consistent with the explanatory direction proposed in the main text, namely

that individuals initially show lower response sensitivity to negative feedback originating from robots and exhibit an "accumulation effect." By contrast, participants' acceptance of feedback from human teammates had already fallen to a relatively low level at the first instance of negative feedback, which also supports the "immediate response" and "high sensitivity" characteristics proposed in the main text.

Under the multiple-round positive-feedback condition, feedback acceptance remained stable overall and was not significantly affected by feedback source, corresponding to the stable pattern of behavioral trust and willingness for future cooperation under positive feedback.

Note: Figure translations are in progress. See original paper for figures.