Real-time Facial Communication Restores Cooperation After Defection in Social Dilemmas

Mayada Oudah[1,*] and John Wooders[1,2]

[1]Social Science Division, New York University Abu Dhabi, UAE.
[2]Center for Behavioral Institutional Design, New York University Abu Dhabi, UAE.

[*]Correspondence should be addressed to mayada@nyu.edu

Abstract

Facial expressions are central to human interaction, yet their role in strategic decision-making has received limited attention. We investigate how real-time facial communication influences cooperation in repeated social dilemmas. In a laboratory experiment, participants play a repeated Prisoner's Dilemma game under two conditions: in one, they observe their counterpart's facial expressions via gender-neutral avatars, and in the other no facial cues are available. Using state-of-the-art biometric technology to capture and display emotions in real-time, we find that facial communication significantly increases overall cooperation and, notably, promotes cooperation following defection. This restorative effect suggests that facial expressions help participants interpret defections less harshly, fostering forgiveness and the resumption of cooperation. While past actions remain the strongest predictor of behavior, our findings highlight the communicative power of facial expressions in shaping strategic outcomes. These results offer practical insights for designing emotionally responsive virtual agents and digital platforms that sustain cooperation in the absence of physical presence.

Introduction

Facial expressions play a critical role in shaping human interactions. As a fundamental channel of nonverbal communication, facial expressions convey emotions and intentions without the need for words, yet significantly influence others' behavior and expectations [1, 2, 3, 4, 5]. They guide social behavior, evoke empathy, and help regulate cooperation and conflict in both casual and strategic settings such as negotiations, team decision making, and everyday interactions [6, 7, 8, 9]. The ability to recognize and interpret facial expressions in such settings allows individuals to better anticipate reactions and adjust their decisions accordingly [10].

Despite its central role, much of the economic literature on strategic decision-making has abstracted away from nonverbal communication [11, 12]. In particular, the canonical Prisoner's Dilemma and its repeated versions typically do not account for facial signals, even though such cues are omnipresent in real-world social dilemmas [13, 14, 15]. Understanding the role of facial expressions in strategic interactions

is therefore not only scientifically relevant but also practically important in a digital world where in-person encounters are increasingly replaced by text-based communication.

Recent advancements in biometric technology now allow us to investigate the importance of facial expressions with precision. In this study, we utilize state-of-the-art technology to automatically and accurately measure and classify facial expressions in real time, via a webcam, as individuals interact. An individual's captured expression is represented to their counterpart by an avatar.

Crucially, this methodology overcomes a major hurdle in behavioral analysis: the signal-to-noise problem inherent in human faces. In standard interactions, the "signal" of an emotion is often confounded by the "noise" of the sender's identity—where attributes like race, gender, and attractiveness can overpower the expression itself and form biases [16]. By re-targeting biometric data to gender-neutral avatars, we effectively isolate the emotional signal from these identity cues . We design and evaluate a novel set of gender-neutral avatars to accurately represent the universal set of facial expressions as defined by Ekman [1].

In prior research on the role of emotions in social dilemmas [17, 18, 19, 20], by contrast, participants select a facial expression or a text description of their emotional state from a predefined set of faces or states, either before or after making a decision. In this case, participants may behave strategically in conveying their emotional state. We address this challenge in our study by capturing and communicating facial expressions in real time as participants interact with each other.

In our experiment, pairs of participants play the repeated Prisoner's Dilemma game for at least 30 rounds, with a continuation probability of 90% thereafter. We study how the ability to observe facial expressions – communicated in real-time via avatars – affects cooperation rates. In the Facial Communication (FC) treatment, players see their counterpart's facial expressions as represented by a gender-neutral avatar that reflects the most intense emotion detected by the biometric software. In the control treatment, no Facial Communication (nFC), players receive no visual emotional feedback.

Our results show that, in addition to sustaining high mutual cooperation rates, the ability to observe facial expressions significantly increases cooperation following defection. This suggests that visual social cues soften the perceived severity of non-cooperative behavior, encouraging participants to respond with greater forgiveness. These findings highlight the power of embedding real-time emotional feedback in social dilemmas, where facial expressions play a restorative role in maintaining cooperative dynamics.

Results

To evaluate the avatars we designed, shown in Figure 1, we invite subjects to participate in a Qualtrics survey. In the survey we ask each participant to classify the avatars into 9 categories, consisting of the 8 universal emotions (Joy, Sadness, Surprise, Fear, Anger, Disgust, Contempt, Neutral) and the category "Hard to Classify." We also ask participants to classify the avatars by gender (Male, Female, Can be Either) by dragging an image of the avatar and dropping it into the category where they believe it belongs. Figure 2 shows the participants' classification accuracy. All of the avatars achieved a classification accuracy rate greater than 80%, with the exception of *Contempt* (accuracy $\approx$ 57%). This lower accuracy is consistent with established literature, which documents that contempt is frequently conflated with Disgust or Anger and is difficult to recognize [21]. Indeed, the majority of misclassifications for this avatar fell into "Anger", "Disgust", and "Hard to Classify" categories (see Supplementary information Figure S1), confirming that the avatar successfully captured the intended negative valence despite the semantic ambiguity. Figure 3 shows the participants' gender classification. For each avatar except Contempt, the majority of participants assigned "Can be Either" as the gender classification.

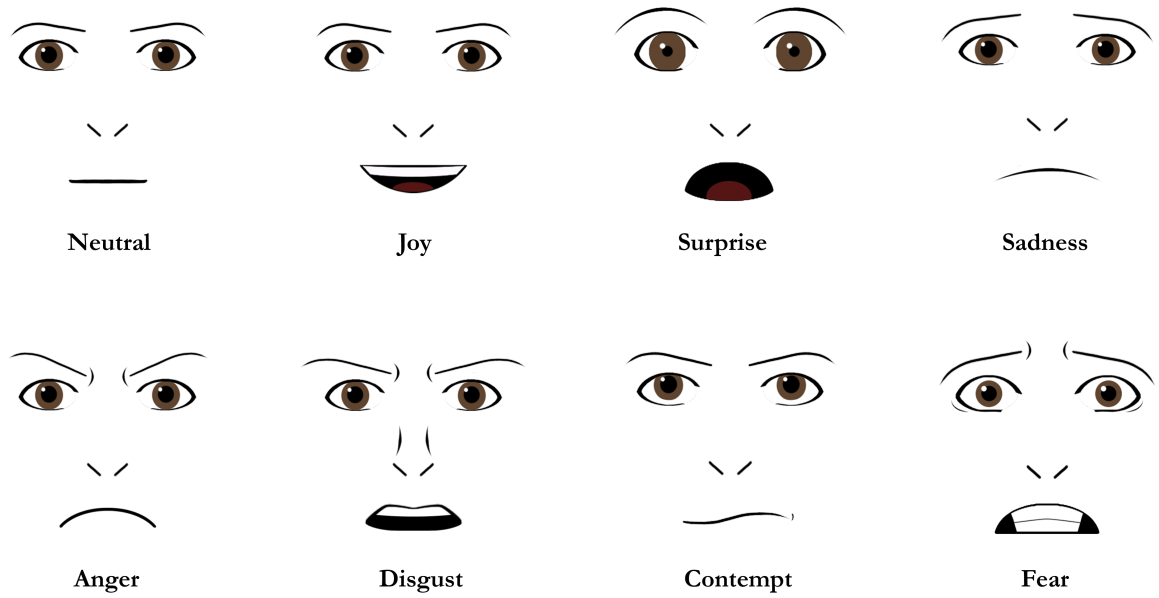

Figure 1: **The avatars used to represent facial expressions.** The avatars utilize a schematic design to minimize the "Uncanny Valley" effect while maximizing signal clarity.

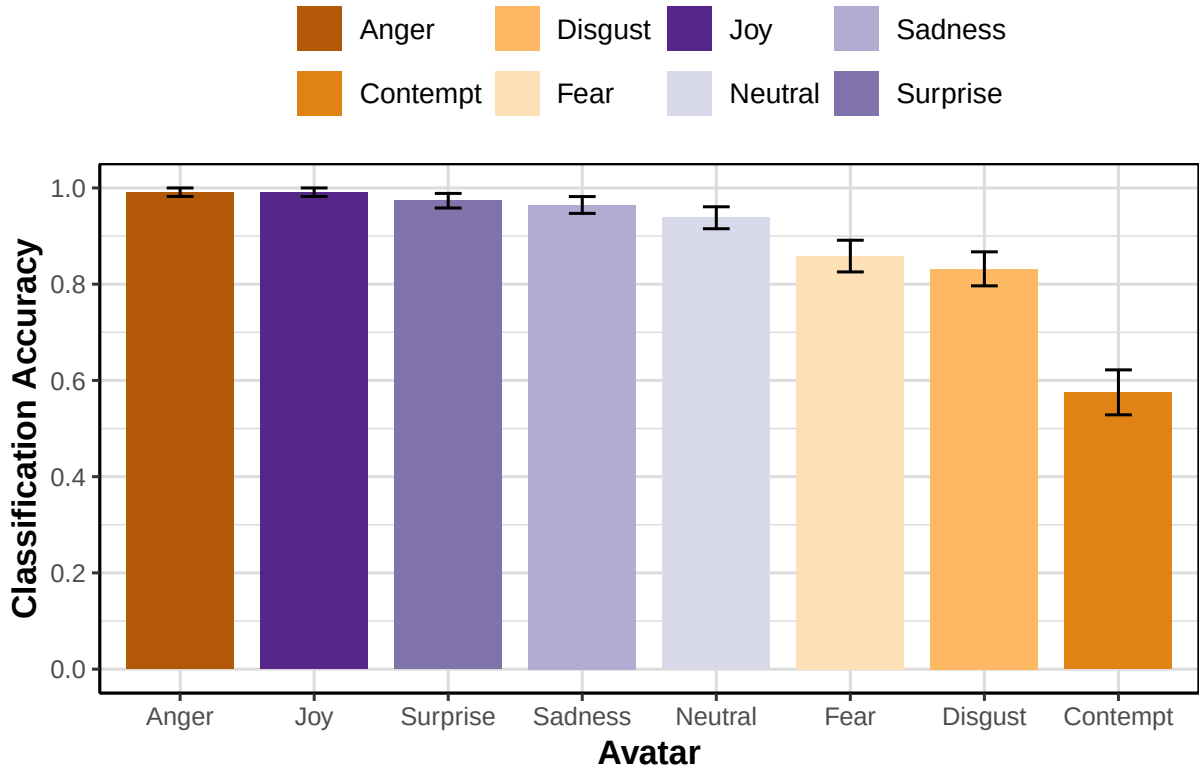

Figure 2: **Avatars classification accuracy.** Error bars show the standard error of the mean.

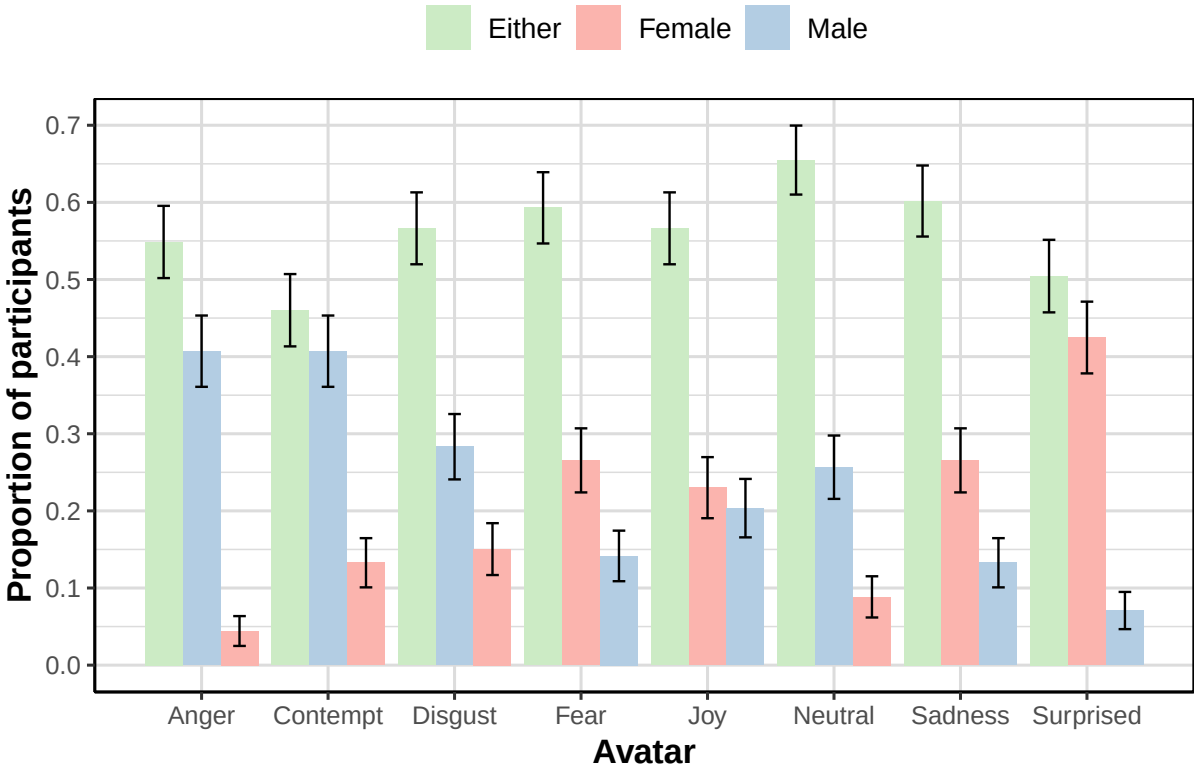

Figure 3: **Avatars gender classification.** Error bars show the standard error of the mean.

Given the high classification accuracy of the avatars (except for Contempt) and that "Can be Either" is the most frequent gender classification, we use these avatars to represent participants' facial expressions in the Prisoner's Dilemma experiment. The experiment has two treatments that differ in whether a player observes the avatar of his counterpart. We label the treatment in which facial communication is enabled as FC and the treatment in which facial communication is disabled as nFC (for further details, see the Methods section). In the FC treatment, facial expressions are communicated via an avatar. Each player observes, in real-time on their own screen, the facial expression of the other player.

Figure 4 shows the cooperation rates for both treatments. Participants cooperate more in the FC treatment than in the nFC treatment. This result is validated through regression analysis, using the Generalized Linear Mixed Model (GLMM) to account for within-subject dependencies and repeated measurement. Figure 5 visualizes the regression coefficients (see Supplementary Table S1 for full model details). The positive coefficient for FC ($\beta = 0.47, p < 0.001$) indicates that the presence of facial avatars significantly increases cooperation rates compared to the anonymous baseline.

To understand the influence of perceived facial expressions on the decision to cooperate, we regress whether or not a participant cooperated on (i) own and other's displayed facial expression prior to making a decision and (ii) the cooperation decisions of both players in the prior round. Figure 6 shows that while the presence of Joy (Own and Other) is a positive predictor of cooperation, recent history of actions (Own and Partner's Previous Cooperation) is the strongest predictor of current behavior, confirming the dominance of reciprocal strategies such as Tit-for-Tat. See Supplementary Table S2 for full model details

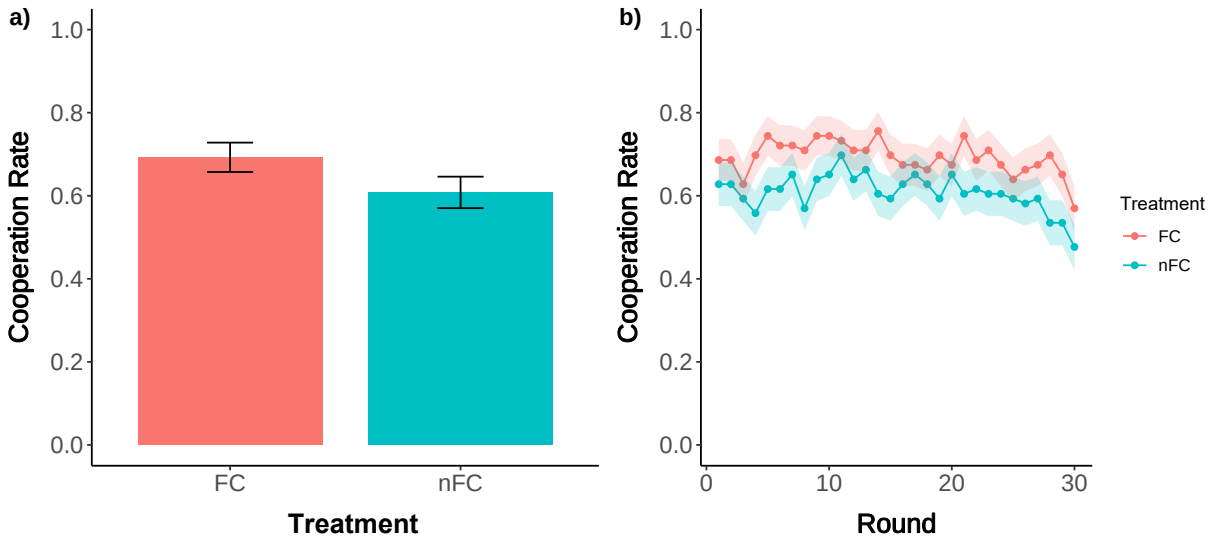

Figure 4: **Experimental results – Cooperation. a)** Shows the cooperation rate in each treatment. Error bars show the standard error of the mean. **b)** Depicts cooperation rates across 30 rounds.

We run further analysis to investigate whether participants' facial expressions are influenced by their own and the counterpart's actions. Crucially, this analysis relies on the raw biometric values captured by the iMotions software, which differs from the facial expressions displayed to the counterpart. Unlike the avatar, which applied a 60% intensity threshold to ensure signal clarity, these measured values capture subtle emotional shifts and internal reactions that may not have been visible to the other player.

Figure 7 shows how participants' own actions influence these internal emotional states. It is evident that participants are more emotionally engaged (expressing more non-Neutral emotions) in the FC treatment than in the nFC treatment. When participants cooperate, they tend to express positive emotions – especially Joy – suggesting emotional satisfaction or prosocial signaling (see Figure 7 (a) and (c)). This aligns with the view that cooperation is emotionally rewarding.

In the FC treatment, own defection is associated with less Joy and a rise in negative emotions (see Figure 7 (b)). This biometric spike potentially indicates internal conflict, guilt, or social disapproval—affective reactions that occur even if the expression was too subtle to trigger the avatar. In contrast, in the nFC treatment, own defection is not associated with any significant change in the defector's measured emotional state, highlighting the lower emotional stakes of anonymous interaction.

Figure 8 illustrates the influence of the counterpart's actions on the participant's measured internal emotional state. Consistent with the findings on own actions, participants in the FC treatment exhibit significantly higher emotional reactivity than those in the nFC treatment.

When the counterpart cooperates, participants respond with a distinct rise in Joy, reflecting social approval, gratitude, or a shared prosocial orientation (see Figure 8 (a) and (c)). Conversely, when the counterpart defects, it triggers strong negative biometric signals in the FC treatment—specifically spikes in Contempt and Sadness (see Figure 8 (b)). This sharp negative reaction (absent in the anonymous condition) suggests that when facial identity is present, defection is perceived not merely as a strategic loss, but as a social or moral violation.

Crucially, the timeline of emotional states reveals a difference in expectations. In the nFC treatment, "pre-Joy" (measured Joy before the outcome is revealed) is significantly lower in rounds where the counterpart eventually defects (see Figure 8 (d)). This indicates that in anonymous play, defection is often the status quo; participants are trapped in a cycle of mutual defection and negative affect, correctly anticipating further betrayal.

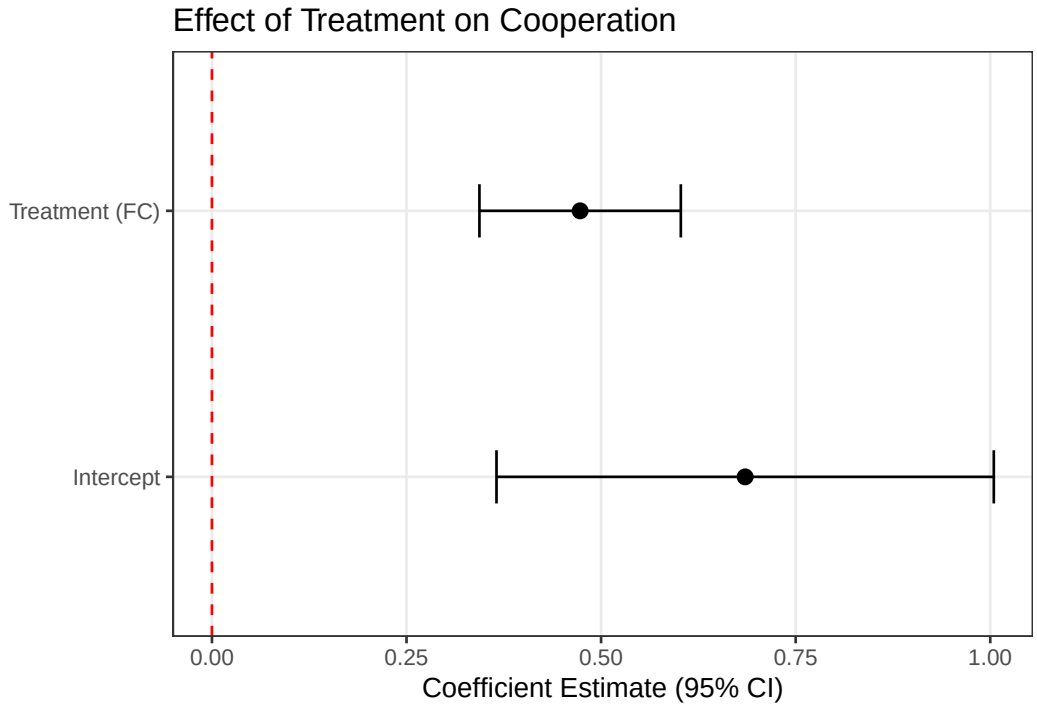

Figure 5: **Facial communication significantly increases the likelihood of cooperation.** Fixed-effect coefficient from a GLMM predicting cooperation ($N = 5,160$). The point estimate represents log-odds relative to the nFC baseline; error bars denote 95% confidence intervals.

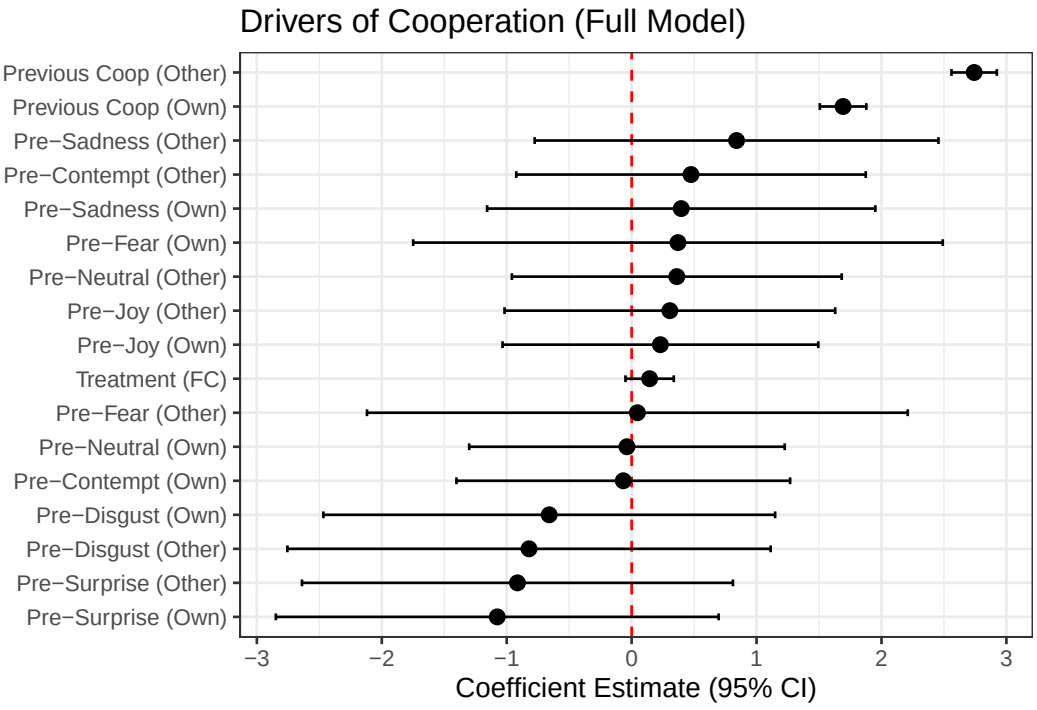

Figure 6: **History of play is the strongest predictor of cooperation.** GLMM estimates comparing structural and emotional variables. Points represent log-odds; error bars denote 95% confidence intervals.

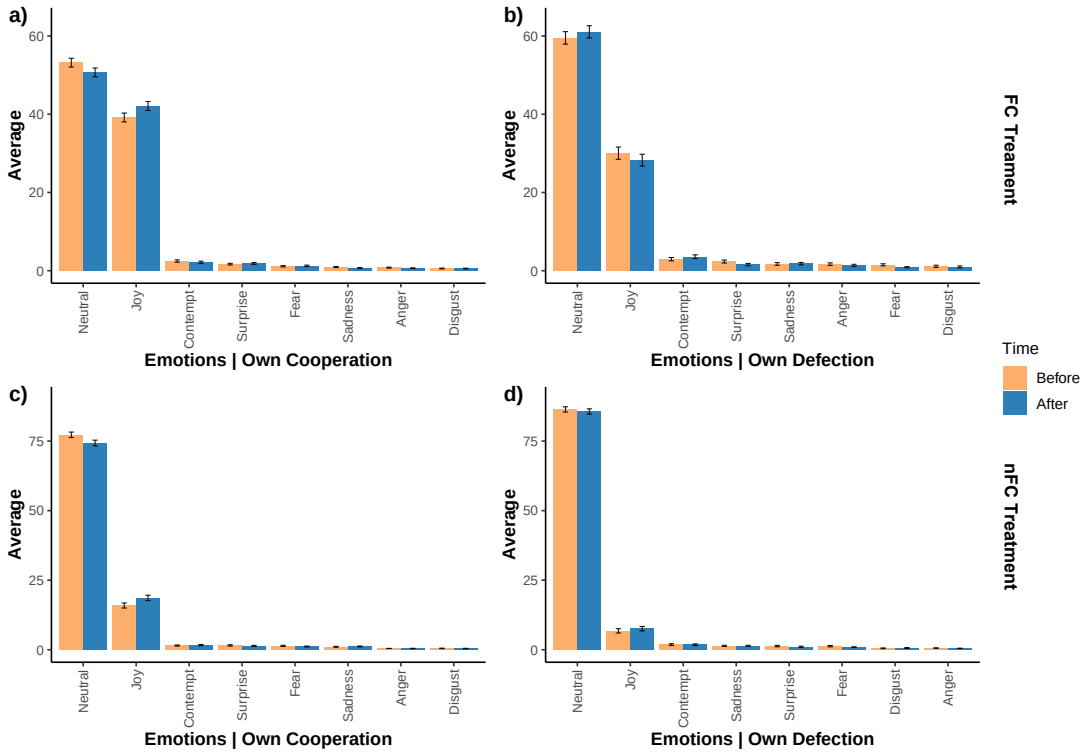

Figure 7: **Measured Emotions Before and After Own Action.** Error bars show the standard error of the mean.

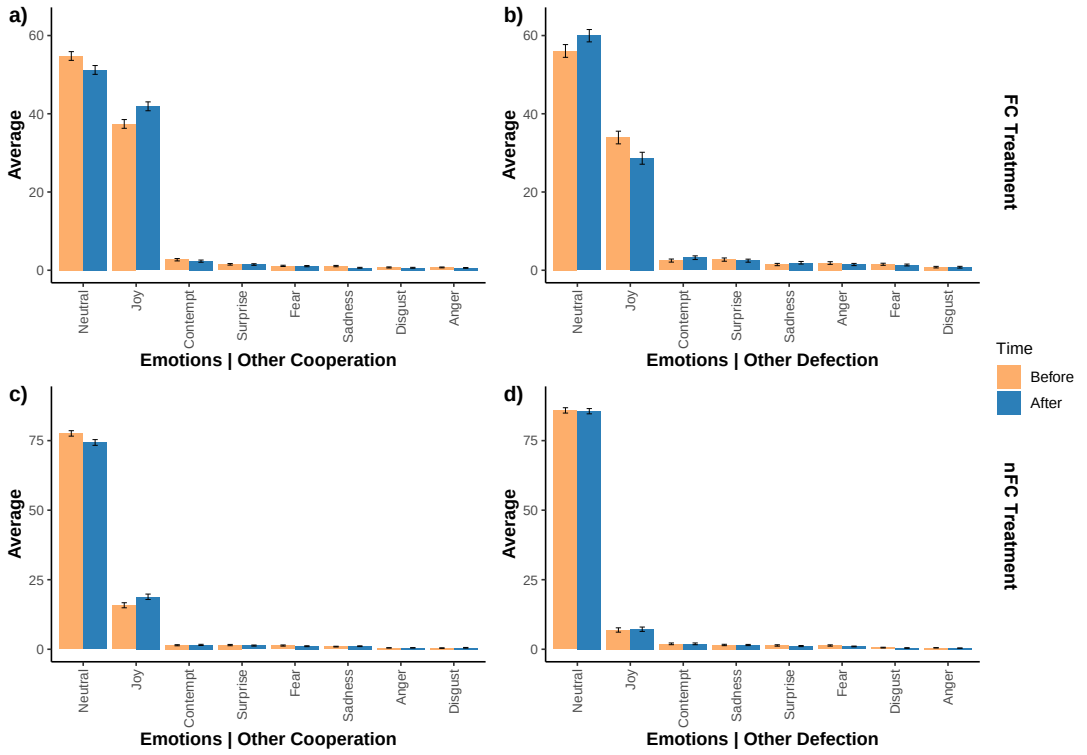

Figure 8: **Measured Emotions Before and After The Other Player's Action.** Error bars show the standard error of the mean.

In contrast, in the FC treatment, pre-Joy remains high even in rounds where the counterpart subsequently defects. This suggests that FC participants maintain a state of emotional optimism or cooperative intent right up until the moment of betrayal. The defection in the FC treatment therefore acts as a violation of this sustained hope, explaining the intense restorative efforts (forgiveness) that follow. Further information can be found in our Supplementary information.

Figure 9 (a) shows the participants' cooperation rate given that they cooperated in the previous round, for both FC and nFC treatments. Likewise, Figure 10 (a) shows the participants' cooperation rate given that their counterpart cooperated in the previous round, for both treatments. In each case, there is no statistically significant difference in cooperation rates between the treatments.

There is, however, a striking difference between the treatments if a player defected in the previous round. Figure 9 (b) shows that participants are significantly more likely to cooperate in the FC than in the nFC treatment following their own defection in the previous round. Participants may feel a sense of guilt and regret following their own defection when facial expressions are communicated between players, that they don't feel in the absence of facial communication. Figure 10 (b) shows that participants are also significantly more likely to cooperate in the FC than in the nFC treatment following their counterparts' defection. In the presence of facial communication, participants are more forgiving and more cooperative after their counterparts' defection. This suggests that visual social cues (facial expressions) play a restorative role in maintaining cooperative dynamics.

At the end of the experiment, each participant completed a survey to rate the other player's likability, intelligence, cooperativeness, trustworthiness, selfishness, forgivingess, tendency to bully, and whether they would like to interact with them again (see Figure 11 (a) and the Supplementary material for more details). We find a statistically significant difference in the reported "Forgivingess" of the other player between the two treatments. FC participants perceive the other player to be more forgiving than nFC participants (see Figure 11 (b)). This result supports that Facial Communication promotes forgiveness following a defection – see Figure 10 (b).

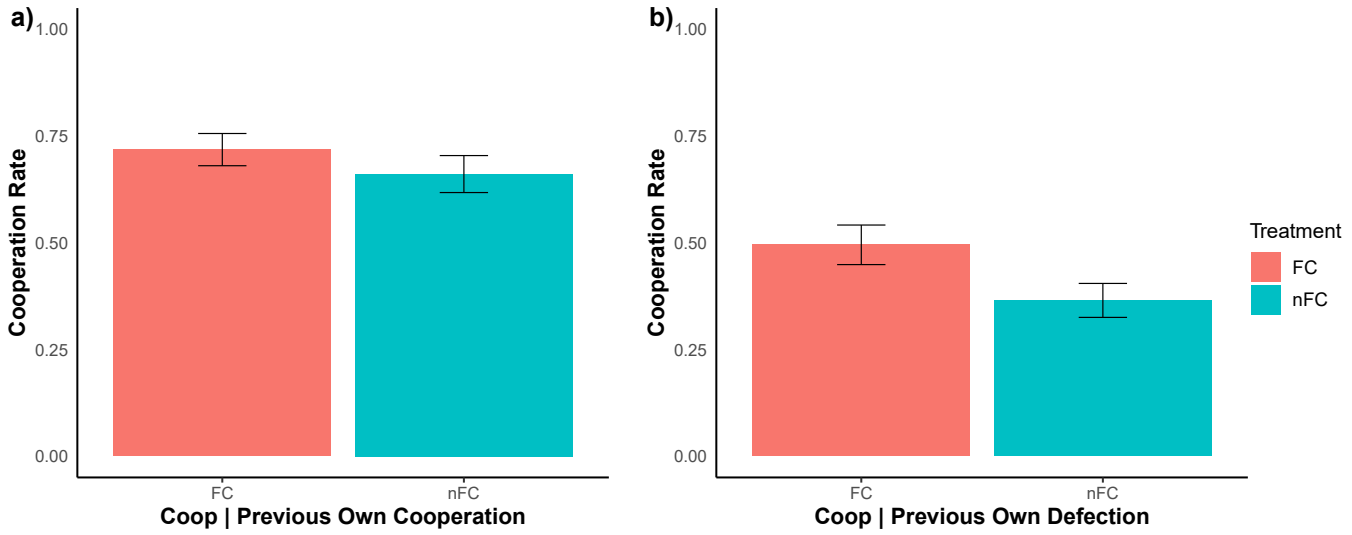

Figure 9: **Conditional Cooperation of Own Previous Decision. a)** Cooperation rate given own cooperation in the previous round and **b)** Cooperation rate given own defection in the previous round. Error bars show the standard error of the mean.

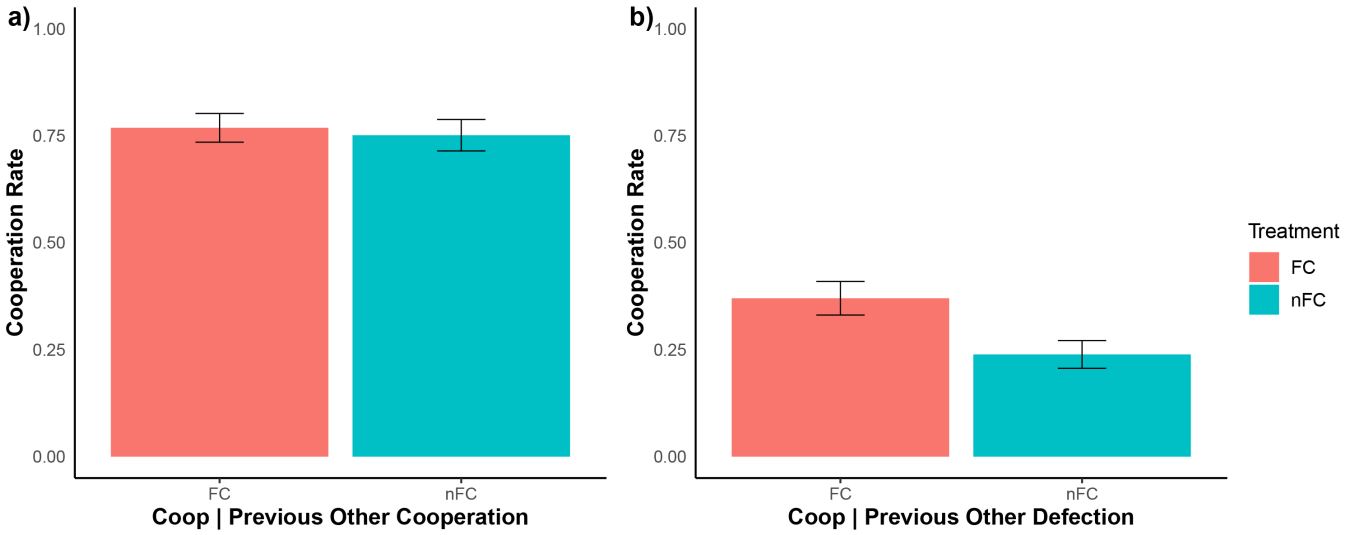

Figure 10: **Conditional Cooperation of the Other Player's Previous Decision.** **a)** Cooperation rate given the other player's cooperation in the previous round and **b)** Cooperation rate given the other player's defection in the previous round. Error bars show the standard error of the mean.

Discussion

Our results show that face-to-face communication via avatars promotes cooperation in social dilemmas, fundamentally changing the perception of a defection. In the nFC treatment, a defection is just a cold numerical choice and the standard tit-for-tat logic ("you defect, I defect") dominates. But when the participant can see their counterpart's avatar before and after their counterpart defects, they obtain extra social and emotional information that lets them handle their betrayal with more forgiveness. That is, in the FC treatment, one might pick up on an expression – e.g., Sadness following defection and Joy following cooperation. Those facial cues signal that the defection may have been unintentional or that the defector feels bad about it, which makes the participant more willing to give them another chance. By contrast, in the nFC treatment, the participants see the outcome without any avatar. There is no socially costly apology-signal to soften the blow, and thus they retaliate more than in the FC treatment.

The effect of facial communication is especially pronounced after a defection, whether one's own or the counterpart's. In standard repeated games, defection typically initiates a downward spiral, with reciprocity mechanisms punishing defections. Here, however, facial expressions appear to soften the rigidity of that response. Players in the FC treatment are more likely to re-establish cooperation after it has been breached, indicating that facial cues provide an interpretive context to actions that would otherwise be perceived as purely self-interested and hostile.

Facial expressions function as affective "repair tools," helping participants to attribute remorse or alternative motives to a defection that might otherwise be viewed as intentional. This is consistent with work showing that emotions – particularly those expressed facially – are not just reflective but communicative; they convey intent, invite empathy, and facilitate reconciliation [4, 5, 19].

While the history of play establishes the baseline for cooperation, our analysis reveals that facial expressions function as a critical 'circuit breaker' specifically when that baseline is fractured. Although past actions remain the primary driver of routine behavior – consistent with standard reciprocity – emotional cues become decisive following a violation. By recontextualizing an unexpected defection, facial expres-

sions modulate the rigid logic of tit-for-tat, allowing players to prioritize forgiveness over retaliation. Thus, while memory sustains stability, emotional communication restores it.

Importantly, players in the FC treatment not only behaved more cooperatively but also were more emotionally engaged. Their facial expressions, especially Joy and Contempt, were more pronounced and responsive to both their own and others' actions. This increased expressiveness suggests that being seen, and seeing others, increases the social salience of the interaction and promotes cooperation.

These findings carry several practical implications. In digital settings, where facial feedback is usually absent, basic avatar-driven emotional expressions can revive elements of accountability, empathy, and shared understanding. For designers and developers of virtual agents and AI systems, our results highlight the importance of embedding real-time emotional feedback loops that are interpretable and responsive, not merely reactive.

Our experimental design involves specific trade-offs inherent to the digital mediation of human affect. While our real-time emotion mapping ensures strict experimental control, the imposition of a 60% intensity threshold likely filters out subtle micro-expressions. By suppressing low-intensity signals (e.g., a faint smile below the threshold), our design prioritizes signal clarity over sensitivity to prevent avatar flickering or signal instability. Consequently, our results may represent a conservative estimate of the restorative power of facial expressions, as subtle conciliatory cues were effectively censored. Despite the artificial nature of the avatars, the high ratings for Likability in Figure 11 (a) show that participants treated the interaction as a genuine social exchange, not just a human-computer interaction task.

All put together, our findings suggest that even subtle facial communication, in controlled conditions, can meaningfully shift behavioral outcomes in strategic environments. By transforming how individuals interpret defection, facial expressions provide a window into intent and a platform for emotional repair. Future research could explore how these mechanisms operate across different cultural contexts, group sizes, or in high-stakes environments, as well as investigate whether artificial agents designed to express human-like emotions can similarly foster trust and cooperation. Understanding how emotional visibility shapes long-run dynamics is especially timely in an increasingly digital world – where the capacity to "see and be seen" may determine not only what we choose, but who we become in relation to others.

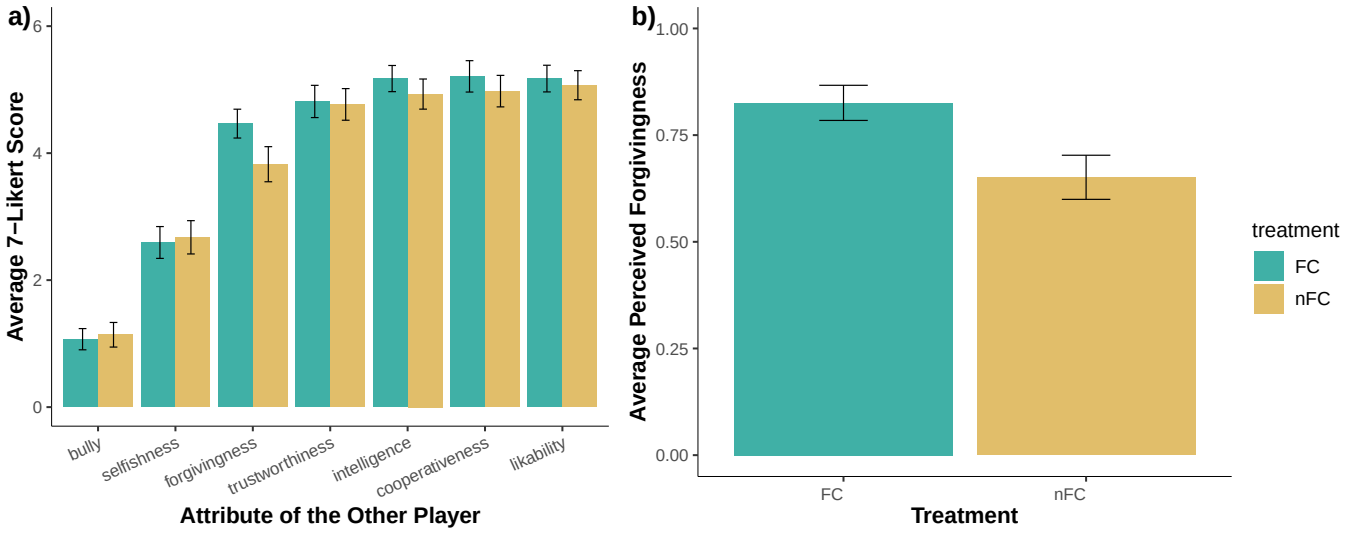

Figure 11: **Participants' perception of the other player's attributes. a)** The average perceived scores for each attribute of the other player. **b)** The frequency that participants report a Likert score of 3 or above of the other player's forgivingess. Error bars show the standard error of the mean.

Methods

All of our protocols received IRB approval from the NYU Abu Dhabi (NYUAD) Internal Review Board. In the first experiment, the subjects assessed the accuracy of the designed avatars. The second experiment, on the interactive Prisoner's Dilemma, followed a between-subject design, and participants were recruited for different treatments on different days. Participants were not aware of the different treatments. Subjects signed a consent form before participating.

Experimental design and setup

To investigate the impact of facial expressions on decision making, we first designed our own set of gender-neutral avatars to represent the universal set of 8 emotions identified by Ekman [1]: neutral, sadness, happiness, fear, anger, surprise, disgust, and contempt.[1] Figure 1 shows the avatars created. The schematic nature of the avatars was a deliberate choice to avoid the "Uncanny Valley" effect, where hyper-realistic but imperfect avatars cause revulsion [22, 23]. Showing avatars instead of the players' actual faces eliminates potential decision biases towards facial characteristics such as gender, ethnicity, attractiveness, and so on [24, 25, 26]. To verify the accuracy of the designed avatars, in the first experiment subjects evaluate the avatars by classifying them according to the universal emotions and according to gender (Male, Female, Can be Either).

We chose to study the Prisoner's Dilemma since it is an engaging game where players might exhibit a broad range of facial expressions. In the experiment, we used the graphical user interface (GUI) of Blake et al. [27] (see Figure 12). In the game, two players face each other, with the participant being the player at the bottom and his counterpart being the player on top. Each player can either "push" or "pull" their tray by clicking the corresponding button. The game is played simultaneously since a player does not observe their counterpart's action before they take their own. The game is played by fixed pairs.

We build on the original GUI by adding a block in the middle of the screen where the counterpart's facial expression is represented via an avatar. In particular, we display the avatar corresponding to the most intense emotion, as measured by the biometric software, so long as its intensity exceeds a threshold of 60%; otherwise we display the Neutral avatar. If the software does not detect a face, the block is empty.

There are two experimental treatments: (i) Facial Communication (FC) and (ii) no Facial Communication (nFC). In both treatments, the participants' facial expressions are captured and measured in real-time using iMotions [28], which provides accurate and automatic facial expression analysis using Affectiva Affdex [29, 30]. The software detects and measures facial expressions via a webcam on top of each monitor. In the FC treatment, participants observe their counterpart's avatar throughout the entire experiment. In the nFC treatment, participants' facial expressions were recorded but not displayed. See the Supplementary material for more details about iMotions platform.

The experimental sessions were held in the Social Science Experimental Lab (SSEL) at NYUAD, where participants sat at private cubicles. We invited participants in groups of 8 per session. After participants read and signed consent forms, the experimenter reviewed the instructions and participants undertook a tutorial on their computer to familiarize themselves with the mechanics of taking an action in the game and calibrate the software to their facial expressions. Participants were informed that their facial expressions are recorded and analyzed throughout the play of the PD game, and were instructed to avoid covering their face. Participants were informed that they would play the PD game with another fixed participant for at least 30 rounds, after which there would be a 10% chance that the interaction would terminate at the end

[1] We designed our own avatars since, in the existing literature, avatars are often gendered.

of each following round. In this setting, cooperation at each round is sustainable as a subgame perfect equilibrium using strategies like Grim Trigger or Tit-for-Tat. Participants were also informed they would be paid for two random rounds at the end of the experiment.

Once the instructions were presented, the experimenter attended to each participant and ensured they were seated properly for the webcam (Logitech C920 Widescreen HD Pro) to accurately capture their face and they were requested to make several facial expressions to ensure that the iMotions module could detect their facial expressions. After the iMotions calibration was complete, the participants were asked to input their player ID (which was provided to them by the experimenter upon entering the lab), their age, and gender in the game software. Once the information was entered by paired players, the game began. iMotions software was running in the background while the game was active. There was no timer set for participants to make a decision in each round. The game lasted approximately 10–15 minutes, depending on the number of rounds realized and how fast participants made decisions. When the interaction ended, a screen with the total earnings was displayed. Participants filled out a post-experiment survey that had questions related to how they made their decisions in the game, demographics, and how do they rate the other player's behavior. Participants were paid using YouGotaGift, a digital gift card company in the Middle East that supports electronic gift cards of over 700 brands, including Amazon.

Figure 12: **User interface and treatments.** **a)** The interface of our 2-player Prisoner's Dilemma game under the FC treatment, where the player can observe the avatar of the other player. Participants control the player on the bottom and see the other player on the opposite side. Each player can either push (Cooperate) or pull (Defect) the lever by clicking the corresponding buttons for these actions. Pushing the lever delivers three coins to the other player and causes one coin to fall into a black hole. Pulling the lever delivers one coin to the player and causes three coins to fall into the black hole. **b)** The interface under the nFC treatment. The rules are the same as the FC treatment, but the player does not observe the avatar of the other player. **c)** The interface of mutual cooperation. **d)** The interface of cooperating against a defector.

Recruitment

We recruited students from NYUAD, who were 18 years or older, using the ORSEE recruitment system to evaluate the designed avatars and to participate in our experimental studies. To be eligible to participate in the PD game, students could not have a thick beard or wear glasses, in order to avoid distorting the quality of the data collected by iMotions. Students who evaluated the designed avatars were also allowed to participate in the PD game if they satisfied both of these eligibility criteria. A student was not allowed to participate in both the FC and the nFC treatment.

Data collection for the survey took place between the 14^{th} of October and the 3^{rd} of November, 2022. For the FC Treatment, it took place between the 29^{th} of May and the 16^{th} of June, 2023, and for the nFC treatment, it took place between the 6^{th} and the 24^{th} of November, 2023. We recruited 113 participants for the avatar evaluation survey, and a total of 172 participants for the PD game, 86 for the FC treatment and 86 for the nFC treatment.

Participants received AED 15 ($4.08) for completing the evaluation survey, which took an average completion time of 8.60 minutes. The participants in the PD game received an AED 30 ($8.17) show-up fee in addition to a bonus that ranged between AED 0 ($0) and AED 190 ($51.73) depending on their decisions and the randomly selected rounds for payment. The average compensation in the FC treatment was AED 121.86 ($33.18) and in the nFC treatment was AED 116.02 ($31.59). The average completion time was 20 minutes.

Subject characteristics

Participants who completed the avatar evaluation survey were 56.64% female and 42.48% male, 36.28% of which come from South Asia (16.81% from Europe and Central Asia, 14.16% from East Asia and Pacific, 11.50% from Middle East and North Africa, 10.62% from Sub-Saharan Africa, 6.19% from Latin America and Caribbean, and 3.54% from North America), with an average age of 20.97 (sd = 1.38). Participants who completed the FC experimental treatment were 60.47% male and 39.53% female, with an average age of 20.31 (sd = 1.51). Participants who completed the nFC experimental treatment were 52.33% male, 46.51% female, and 1.16% other. The participants' average age was 20.27 (sd = 1.46).

Data availability

All data analyzed in these studies will be deposited on Open Science Framework at the time of publication and will be made available for review upon request.

Code availability

The plots, regressions, and statistical analysis were executed using R (version: 4.3.2). The code will be deposited on Open Science Framework at the time of publication and will be made available for review upon request.

References

[1] Ekman, P. An argument for basic emotions. Cognition and Emotion **6**, 169–200 (1992).

[2] Frith, C. Role of facial expressions in social interactions. Philosophical Transactions of The Royal Society Biological Sciences **364**, 3453–3458 (2009).

[3] Knapp, M. L., Hall, J. A. & Horgan, T. G. Nonverbal Communication in Human Interaction (8th ed.) (Boston, MA: Cengage Learning, 2013).

[4] Sagliano, L., Ponari, M., Conson, M. & Trojano, L. Editorial: The interpersonal effects of emotions: The influence of facial expressions on social interactions. Frontiers in Psychology **13** (2022).

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

[6] Hatfield, E., Cacioppo, J. T. & Rapson, R. L. Emotional contagion. Current Directions in Psychological Science **2**, 96–99 (1993).

[7] Holland, A. C., O'Connell, G. & Dziobek, I. Facial mimicry, empathy, and emotion recognition: a meta-analysis of correlations. Cognition and Emotion **35**, 150–168 (2021).

[8] Kleef, G. A. V. How emotions regulate social life: The emotions as social information (easi) model. Current Directions in Psychological Science **18**, 184–188 (2009).

[9] Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M. & Seabright, P. Honest signaling in trust interactions: smiles rated as genuine induce trust and signal higher earning opportunities. Evolution and Human Behavior **36**, 8–16 (2015).

[10] Ekman, P. Emotions revealed: Recognizing faces and feelings to improve communication and emotional life (Times Books/Henry Holt and Co., 2003).

[11] Sally, D. A general theory of sympathy, mind-reading, and social interaction, with an application to the prisoners' dilemma. Social Science Information **39**, 567–634 (2000).

[12] Camerer, C. F. Behavioral Game Theory: Experiments in Strategic Interaction (Princeton University Press, Princeton, NJ, 2003).

[13] Axelrod, R. The Evolution of Cooperation (Basic Books, New York, 1984).

[14] Fehr, E. & Gächter, S. Fairness and retaliation: The economics of reciprocity. Journal of Economic Perspectives **14**, 159–181 (2000).

[15] Binmore, K. Playing for Real: A Text on Game Theory (Oxford University Press, Oxford, 2007).

[16] Stirrat, M. & Perrett, D. Valid facial cues to cooperation and trust: Male facial width and trustworthiness. Psychological Science **21**, 349–354 (2010).

[17] Hopfensitz, A. & Reuben, E. The Importance of Emotions for the Effectiveness of Social Punishment. The Economic Journal **119**, 1534–1559 (2009).

[18] Mussel, P., Göritz, A. S. & Hewig, J. The value of a smile: Facial expression affects ultimatum-game responses. Judgment and Decision Making **8**, 381–385 (2013).

[19] de Melo, C. M., Gratch, J. & Carnevale, P. J. Humans versus computers: Impact of emotion expressions on people's decision making. IEEE Transactions on Affective Computing **6**, 127–136 (2015).

[20] de Melo, C. M. & Terada, K. The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. Scientific Reports **10** (2020).

[21] Matsumoto, D. & Ekman, P. The relationship among expressions, labels, and descriptions of contempt. Journal of personality and social psychology **87**, 529–40 (2004).

[22] Mori, M., MacDorman, K. F. & Kageki, N. The uncanny valley [from the field]. IEEE Robotics Automation Magazine **19**, 98–100 (2012).

[23] Oudah, M., Makovi, K., Gray, K., Battu, B. & Rahan, T. Perception of experience influences altruism and perception of agency influences trust in human–machine interactions. Scientific Reports **14** (2024).

[24] Wilson, R. K. & Eckel, C. C. Judging a book by its cover: Beauty and expectations in the trust game. Political Research Quarterly **59**, 189–202 (2006).

[25] Stirrat, M. & Perrett, D. Valid facial cues to cooperation and trust: Male facial width and trustworthiness. Psychological Science **21**, 349–354 (2010).

[26] Chen, J. et al. Electrophysiological correlates of processing facial attractiveness and its influence on cooperative behavior. Neuroscience Letters **517**, 65–70 (2012).

[27] Blake, P. R., Rand, D. G., Tingley, D. & Warneken, F. The shadow of the future promotes cooperation in a repeated prisoner's dilemma for children. Scientific Reports **5** (2015).

[28] iMotions. Facial expression analysis - emotion detection software (2025). URL `https://imotions.com/affectiva/`.

[29] Stöckli, S., Schulte-Mecklenbeck, M., Borer, S. & Samson, A. C. Facial expression analysis with affdex and facet: A validation study. Behavior Research Methods **50**, 1446–1460 (2018).

[30] Kulke, L., Feyerabend, D. & Schacht, A. A comparison of the affectiva imotions facial expression analysis software with emg for identifying facial expressions of emotion. Frontiers in Psychology **11** (2020).

Acknowledgments

The authors are deeply grateful to Professor Rebecca Morton (1954–2020) for her early and foundational contribution to the conception of this study. Although she passed away before the final study was implemented, her ideas and guidance were pivotal in shaping its early direction.

This research is supported by ASPIRE Award for Research Excellence (AARE-2019-48). Wooders gratefully acknowledges financial support from Tamkeen under the NYU Abu Dhabi Research Institute Award CG005.

Author contributions statement

M.O. conceived the study. M.O. and J.W. designed the experiments. M.O. conducted the experiments and collected the data. M.O. and J.W. analyzed the data, interpreted the results, and wrote the manuscript.

Competing interests statement

The authors declare no competing interests.