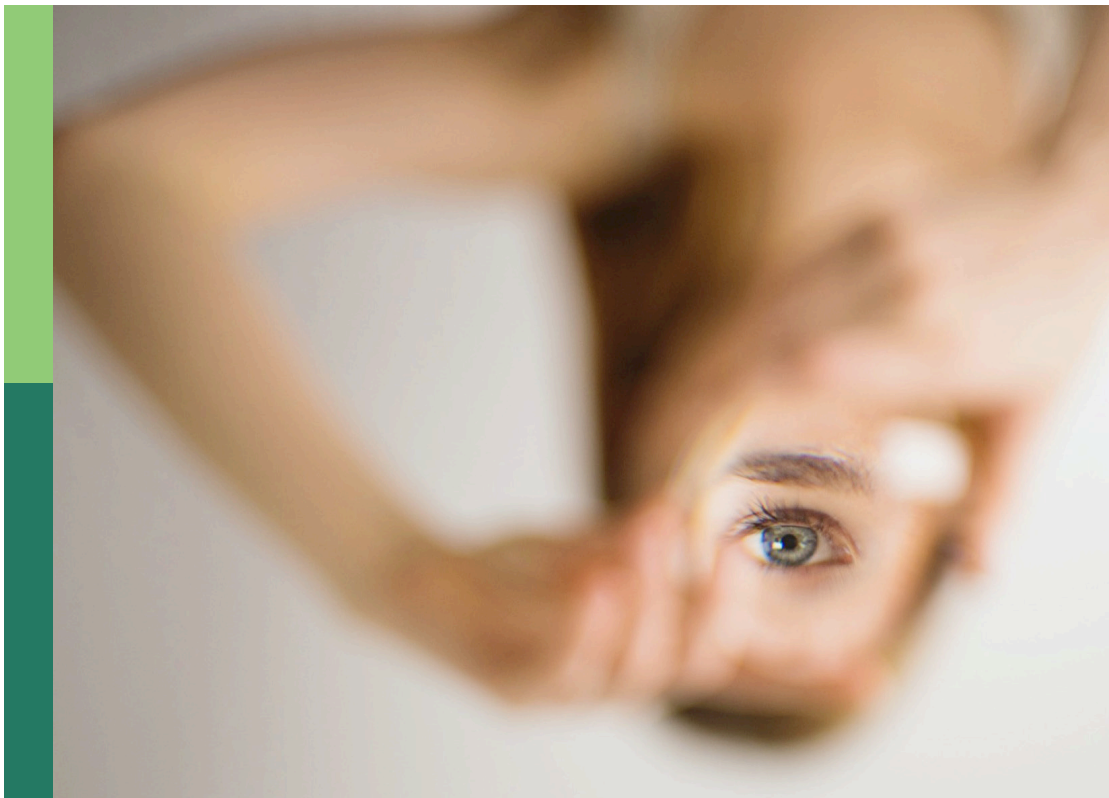
frontiers | Research Topics

The social brain: new insights from social, clinical, and biological psychology

Edited by
Carmelo Mario Vicario, Gabriella Martino, Giuseppe Craparo, Chiara Lucifora and Paola Magnano

Published in
Frontiers in Psychology



ISSN 1664-8714
ISBN 978-2-8325-5992-5
DOI 10.3389/978-2-8325-5992-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

The social brain: new insights from social, clinical, and biological psychology

Topic editors

Carmelo Mario Vicario — University of Messina, Italy
Gabriella Martino — University of Messina, Italy
Giuseppe Craparo — Kore University of Enna, Italy
Chiara Lucifora — University of Bologna, Italy
Paola Magnano — Kore University of Enna, Italy

Citation

Vicario, C. M., Martino, G., Craparo, G., Lucifora, C., Magnano, P., eds. (2025). *The social brain: new insights from social, clinical, and biological psychology*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5992-5

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 29 January 2025
DOI 10.3389/fpsyg.2025.1552456

OPEN ACCESS

EDITED AND REVIEWED BY
Antonino Vallesi,
University of Padua, Italy

*CORRESPONDENCE
Carmelo M. Vicario
✉ cvicario@unime.it

RECEIVED 28 December 2024
ACCEPTED 13 January 2025
PUBLISHED 29 January 2025

CITATION
Vicario CM, Lucifora C, Craparo G, Magnano P and Martino G (2025) Editorial: The social brain: new insights from social, clinical, and biological psychology.
Front. Psychol. 16:1552456.
doi: 10.3389/fpsyg.2025.1552456

Editorial: The social brain: new insights from social, clinical, and biological psychology

Carmelo M. Vicario[1]*, Chiara Lucifora[2], Giuseppe Craparo[3], Paola Magnano[3] and Gabriella Martino[4]

[1]Dipartimento di Scienze Cognitive, Psicologiche, Pedagogiche e Degli Studi Culturali, Università di Messina, Messina, Italy, [2]Dipartimento di Filosofia e Comunicazione, Università di Bologna, Bologna, Italy, [3]Faculty of Human and Social Sciences, Kore University of Enna, Cittadella Universitaria, Enna, Italy, [4]Department of Clinical and Experimental Medicine, University of Messina, Messina, Italy

KEYWORDS

social brain, clinical psychology, biological psychology, social psychology, neuroscience

Editorial on the Research Topic
The social brain: new insights from social, clinical, and biological psychology

The concept of the "social brain" encapsulates the intricate interplay between neural processes and social behaviors, providing a framework for understanding how we navigate our social world. This is suggested by the literature on brain disorders (Vicario and Lucifora, 2021; D'Amico et al., 2024), as well as research on healthy individuals that highlights the role of personality traits (Rinella et al., 2019; Massimino et al., 2019) and coping strategies (Massimino et al., 2024; Rinella et al., 2017).

The recent collection of articles published in *Frontiers in Psychology* under the Research Topic "*The Social Brain: New Insights from Social, Clinical, and Biological Psychology*" presents a diverse range of studies that deepen our understanding of this multifaceted domain.

Zhang, Cai, et al. examined the impact of moral judgment on bystanders' interpersonal trust, identifying trustworthiness as a crucial mediating factor. This research underscores how moral evaluations can significantly influence social relationships and community interactions. Notably, Vicario et al. (2018) demonstrated that hunger and satiety can affect the judgment of ethical violations, suggesting that physiological states can shape moral reasoning and social cognition. Xu et al. investigated the relationship between cognitive reappraisal, empathy, and prosocial behavior in adolescents. Their findings highlight the importance of emotional regulation strategies in fostering empathetic responses, which are essential for constructive social interactions. This is in line with the earlier work by Vicario et al. (2023) providing evidence of altered fear extinction learning in individuals with high vaccine hesitancy during the COVID-19 pandemic, which underscores the significant role that anxiety and emotional regulation play in public health decisions and social behavior. Complementing this, Troncoso et al. focused on the dynamics of empathy by examining sensorimotor and physiological responses during synchronous experiences of suffering, contributing to our understanding of embodied empathy in social contexts.

Zhang, Deng, et al. conducted a systematic review analyzing trustworthiness studies using the Web of Science database, providing critical insights into how trust and social perception are studied across various disciplines. A further systematic

review/meta-analysis was conducted by Dai et al. investigating neural mechanisms of different types of envy.

Guingrich and Graziano explored the implications of attributing consciousness to artificial intelligence, revealing how human-AI interactions can influence subsequent human-to-human interactions. This study offers a novel perspective on the evolving landscape of social cognition in the digital age.

Li et al. examined cognitive control mechanisms involved in honesty and dishonesty across various conflict scenarios, shedding light on the cognitive processes that govern ethical decision-making. In a related study, Lucifora et al. (2021) demonstrated how self-control predicts moral decision-making, showing that individuals with higher self-control exhibit greater ethical considerations in their choices. This body of work, along with Myznikov et al., which investigated the relationship between dark triad personality traits and structural brain changes, indicates a neurobiological basis for personality in shaping social behavior and ethical judgments.

In conclusion, Armas-Vargas et al. focused on the psychometric properties of the CEMA-A questionnaire, assessing motives for lying, which is crucial for addressing ethical behavior and trust in social interactions. Lastly, Zhang, Li, et al. examined how group membership influences adolescents' third-party punishment behaviors, emphasizing the importance of group identity in moral judgment and social dynamics.

The collection of articles featured in this Research Topic serves as a starting point to the strides being made in our understanding of the social brain. By elucidating the interplay of biological, social, and clinical factors, this collection of studies provides a comprehensive framework that advances scientific knowledge and translates into practical strategies for improving social cognition and connectedness in diverse populations. As we continue to explore the intricate matrices of the social brain, the insights gleaned from this body of work are essential for shaping the future of psychological research and practice.

Author contributions

CV: Writing – original draft, Writing – review & editing. CL: Writing – original draft, Writing – review & editing. GC: Writing – original draft, Writing – review & editing. PM: Writing – original draft, Writing – review & editing. GM: Writing – original draft, Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

D'Amico, G., Pasinato, M., and Prior, M. (2024). "Neuropsychological consequences, emotions and length of alcohol abuse: a preliminary study," in *Preliminary Reports and Negative Results in Life Science and Humanities, Vol. 1*. doi: 10.13129/3035-062X/prnr-4218

Lucifora, C., Martino, G., Curcuruto, A., Salehinejad, M. A., and Vicario, C. M. (2021). How self-control predicts moral decision making: an exploratory study on healthy participants. *Int. J. Environ. Res. Public Health* 18:3840. doi: 10.3390/ijerph18073840

Massimino, S., Rinella, S., Buscemi, A., Similia, E., Perciavalle, V., Perciavalle, V., et al. (2019). Digit ratio, personality and emotions in skydivers. *Biomed Rep*. 10, 39–46. doi: 10.3892/br.2018.1174

Massimino, S., Rinella, S., Famoso, S., and Tortorici, M. M. (2024). Coping strategies, personological and affective characteristics. Reciprocity in the oncology patient-caregiver dyad. *Reti Saperi Linguaggi* 1, 199–219. doi: 10.12832/113134

Rinella, S., Buscemi, A., Massimino, S., Perciavalle, V., Tortorici, M. M., Tomaselli, D. G., et al. (2019). Risk-taking behavior, the second-to-fourth digit ratio and psychological features in a sample of cavers. *PeerJ*. 7:e8029. doi: 10.7717/peerj.8029

Rinella, S., Romeo, R., Di Corrado, D., and Massimino, S. (2017). Attentional processes and affective regulation ability of women practicing Hatha Yoga. *Acta Med. Mediterr*. 6, 1067–1072. doi: 10.19193/0393-6384_2017_6_169

Vicario, C. M., Kuran, K. A., Rogers, R., and Rafal, R. D. (2018). The effect of hunger and satiety in the judgment of ethical violations. *Brain Cogn*. 125, 32–36. doi: 10.1016/j.bandc.2018.05.003

Vicario, C. M., and Lucifora, C. (2021). Neuroethics: what the study of brain disorders can tell about moral behavior. *AIMS Neurosci*. 8, 543–547. doi: 10.3934/Neuroscience.2021029

Vicario, C. M., Makris, S., Culicetto, L., Lucifora, C., Falzone, A., Martino, G., et al. (2023). Evidence of altered fear extinction learning in individuals with high vaccine hesitancy during COVID-19 pandemic. *Clin. Neuropsychiatry* 20, 364–369. doi: 10.36131/cnfioritieditore20230417

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 22 September 2023
DOI 10.3389/fpsyg.2023.1238308

OPEN ACCESS

EDITED BY
Carmelo Mario Vicario,
University of Messina, Italy

REVIEWED BY
Elena Shmeleva,
Russian State Social University, Russia
Kui Yi,
East China Jiaotong University, China

*CORRESPONDENCE
Delin Yu
✉ yu.delin@foxmail.com

RECEIVED 11 June 2023
ACCEPTED 06 September 2023
PUBLISHED 22 September 2023

CITATION
Xu Y, Chen S, Su X and Yu D (2023) Cognitive reappraisal and empathy chain-mediate the association between relative deprivation and prosocial behavior in adolescents.
Front. Psychol. 14:1238308.
doi: 10.3389/fpsyg.2023.1238308

Cognitive reappraisal and empathy chain-mediate the association between relative deprivation and prosocial behavior in adolescents

Yanfeng Xu[1], Sishi Chen[1], Xiaojie Su[1,2] and Delin Yu[1]*

[1]School of Psychology, Fujian Normal University, Fuzhou, Fujian, China, [2]Normal College, Urumqi Vocational University, Urumqi, Xinjiang, China

Background: Relative deprivation is one of the factors that influences the development of personality and behavior. However, it is still unclear whether and how relative deprivation decreases the prosocial behavior in adolescents. This study aimed to examine the association between relative deprivation and adolescent prosocial behavior and the role of emotion regulation strategies and empathy in modifying this association.

Methods: The present study included 609 secondary school students (M = 15.42 years, SD = 0.653) in Fujian Province, China. All participants completed the Relative Deprivation Questionnaire, Emotion Regulation Scale, the Basic Empathy Scale, and Prosocial Behavior Scale. The collected data were analyzed using SPSS 25.0 and Mplus 7.4.

Results: Relative deprivation was negatively correlated with cognitive reappraisal, but positively correlated with expressive suppression. Cognitive reappraisal was positively correlated with empathy and prosocial behavior, but expressive suppression was not. Empathy was positively correlated with prosocial behavior. Relative deprivation decreased prosocial behavior through (a) cognitive reappraisal, (b) empathy, and (c) chain mediation of cognitive reappraisal and empathy. No significant mediating effect of expressive suppression was found.

Conclusion: The results indicate that relative deprivation decreases adolescent prosocial behavior, and that cognitive reappraisal and empathy are the potential psychological mechanisms that affect the association between relative deprivation and adolescent prosocial behavior.

KEYWORDS

relative deprivation, cognitive reappraisal, expressive suppression, empathy, prosocial behavior

1. Introduction

Prosocial behaviors are behaviors that are beneficial to others, society, and the nation (Pfattheicher et al., 2022). At the individual level, while one's prosocial behavior is beneficial to others in difficult situations, it also has potential benefits for oneself, such as increased probability of being assisted (El Mallah, 2020), good community reputation (Berman and Silver, 2022), and a sense of meaning and value in life (Klein, 2017); at the societal level, prosocial behavior contributes to a well-functioning and harmonious society (Carlo and Pierotti, 2020).

Given the importance of prosocial behavior, it is considered an essential aspect of social and moral development during adolescence (Hart and Carlo, 2005). Adolescent development is often regarded as a period of social sensitivities that shapes the behavior and character traits of individuals. Therefore, it is necessary to understand the factors that influence the development of adolescent prosocial behavior.

With increased social and economic instability, widening economic disparities, and the rapid growth of online social media and networks, people today may be more likely than ever to make social comparisons and experience feelings of relative deprivation (Power et al., 2020). Relative deprivation refers to the perception of being worse off compared to a certain standard, accompanied by feelings of anger and resentment (Smith et al., 2012). While previous studies have identified that relative deprivation as a significant factor reduces prosocial behavior (Zhang et al., 2016; Pak and Babiarz, 2023; Zhang et al., 2023), but most of the studies mainly focused on adults; there is a lack of research on the adolescents, and the mechanisms underlying the association between relative deprivation and prosocial behavior is still unclear. Adolescence is a special stage of rapid development of self-identity and self-awareness. It is also the period when individuals are susceptible to developing a sense of relative deprivation because of social comparison (Orben et al., 2020). Exploring the pathways through which relative deprivation affects adolescent prosocial behavior would help school mental health teachers implement effective interventions. Thus, the main question addressed in the present study is if and how relative deprivation decreases the prosocial behavior among adolescents.

2. Literature review

2.1. Relative deprivation and prosocial behavior

Previous research has shown that individuals with relative deprivation often tend to reject prosocial behavior (Pak and Babiarz, 2023). For example, when individuals experience relative deprivation caused by unfair social distribution, they show reduced generosity in dictatorial games (Gheorghiu et al., 2021), and the priming of one's relative deprivation reduced the meaningfulness of engaging in prosocial behavior (Zhang et al., 2023). This may be due to relative deprivation highlights individuals' self-perception as victims of injustice and directs their attention to perceived disadvantages (Callan et al., 2017). According to the social information processing model, before engaging in prosocial behavior, individuals first assess their situation, for instance, whether their needs are being met and whether they have sufficient capacity to engage. The preliminary judgment would determine whether they ultimately engage in prosocial behavior (Nelson and Crick, 1999). Individuals suffering relative deprivation, usually develop a cynical view of society and become more self-centered and less concerned about the plight of others (Zitek et al., 2010). Zhang et al. (2016) validated this opinion, suggesting that the tendency to prioritize self-interest over others mediated the effect of relative deprivation on prosocial behavior. Based on the consistent findings from previous studies involving adults, the present study proposes hypothesis H1: Relative deprivation negatively predicts adolescent prosocial behavior.

2.2. Emotion regulation strategies as a mediator

Relative deprivation results in anxiety and attention bias toward a threat (Zhang et al., 2021), which may indicate that individuals with high levels of relative deprivation are more inclined to engage in automatic negative thinking and develop avoidance attitudes when solving emotional problems (Nadler et al., 2020). The theory of emotion suppression proposed by Langner et al. (2012) may explain this phenomenon. When individuals with low social status subjectively perceive themselves to be in a worse position than others, they usually hide or suppress their negative feelings and behaviors to avoid showing their dissatisfaction and anger in the presence of others with high status. Consistently, Liu et al. (2021) also found that individuals with high levels of relative deprivation tend to use expressive suppression strategies to moderate the emergence of negative emotions in more situations, and rarely use cognitive reappraisal strategies.

The use of emotion regulation strategies could predict prosocial behavior. Previous research has revealed that maladaptive emotion regulation strategies (e.g., expressive suppression) are negatively associated with prosocial behavior (Lockwood et al., 2014), whereas the use of adaptive emotion regulation (e.g., cognitive reappraisal) is positively associated with prosocial behavior (Hodge et al., 2023). This phenomenon also occurs in adolescents; Li et al. (2021) found that adolescents who applied cognitive reappraisal tended to be more likely to engage in prosocial behavior than those who applied expressive suppression. This finding could be attributed to the fact that compared to adaptive emotion regulation strategies, maladaptive strategies distract the person and obscure important information on social interactions to the detriment of prosocial behavior (Shaver et al., 2008).

Taken together, it could be inferred that the use of emotion regulation strategies plays a crucial role in the association between relative deprivation and prosocial behavior. However, this opinion remains to be confirmed. In the present study, we focus on two widely studied emotion regulation strategies: expressive suppression and cognitive reappraisal. Previous studies have shown that the two emotion regulation strategies operate in contrasting effects (Zhou et al., 2023), and differ in terms of psychological processes and physiological mechanisms (Bebko et al., 2011; Hermann et al., 2014). Considering with these findings and referencing relevant literature (Zhang et al., 2023), the present study explored the mediating effect of cognitive reappraisal and/or expressive suppression, respectively. Therefore, the present study proposes the following hypothesis H2 and H3: Cognitive reappraisal mediates the association between relative deprivation and adolescent prosocial behavior (H2), and expressive suppression mediates the association between relative deprivation and adolescent prosocial behavior (H3).

2.3. Empathy as a mediator

Many studies have demonstrated the association between feelings of relative deprivation and counter-empathy (e.g., envy and schadenfreude; Leach and Spears, 2009; Neufeld and Johnson, 2016; Zhao and Zhang, 2022). Relative deprivation arises from perceived inequality and can easily lead to anger and resentment—the core components of malicious envy (Lange and Crusius, 2015). According to the deservingness theory (Feather, 1999), when people become

aware of their disadvantaged status, their dissatisfaction with injustice might cause them to feel schadenfreude, especially if they see those in a superior position as undeserving (Feather and Nairn, 2005). Fu et al. (2017) argued that there may be a common psychological mechanism underlying empathy and counter-empathy. Although there is an unproven association between relative deprivation and empathy, based on the demonstrated association between relative deprivation and counter-empathy, it can be theorized that relative deprivation would negatively predict empathy.

The empathy-altruism hypothesis emphasizes that empathy is a direct cause of prosocial behavior (Batson, 2017). Specifically, when individuals feel empathy toward someone in need, they are motivated to take action to help alleviate their suffering, without any expectation of receiving something in return. The results from various types of previous studies have confirmed the substantial association between empathy and prosocial behavior (Davis, 2015). For example, Tang (2015) found that empathy training with primary school students led to a significant increase in the frequency of prosocial behavior. Furthermore, a longitudinal study with adolescents conducted by Wang and Wu (2020) showed that T1 trait empathy competence significantly predicted T2 prosocial behavior. Recently, a meta-analysis including 62 studies revealed an above moderate-strength positive correlation between empathy and prosocial behavior (Yin and Wang, 2023). The evidence from these studies supports the empathy-altruism hypothesis, namely that empathy is an important motivating factor for prosocial behavior.

Combining with the above, adolescents with high levels of relative deprivation may lack the motivation to engage in prosocial behavior owing to low levels of empathy. Hence, the present study proposes hypothesis H4: empathy mediates the association between relative deprivation and adolescent prosocial behavior.

2.4. A chain-mediation model

Decety and Michalska (2010) argue that emotion regulation is an important component of the empathy process. The ability of well-regulated individuals to regulate negative emotions and maintain optimal levels of emotional arousal allows them to increase their attention to situations faced by others. A close association between emotion regulation and empathy has been identified in the previous literature, e.g., Ornaghi et al. (2020) confirmed that emotion regulation skills play an important role in promoting empathy development in children. Thompson et al. (2019) proposed an integrative account of empathy and emotion regulation—this theory suggests effective regulation of one's own emotions can facilitate empathy for others by enabling individuals to better understand and respond to the emotions of others. For instance, if someone could regulate their own feelings of anger well, they may be more capable of empathizing with someone else's anger and provide more compassionate support. Consistent with this theory, Benita et al. (2017) found the ability to emotion regulation promotes prosocial behavior through the mediation of empathy.

For adolescents, cognitive reappraisal is thought to have a positive emotion-regulation effect (Compas et al., 2017), and, thus, promotes empathy and further increases prosocial behavior. Conversely, expressive suppression is thought to have a negative emotion-regulation effect (Schäfer et al., 2017), and, thus, reduces empathy and further decreases prosocial behavior. Since the use of the two emotion regulation strategies is influenced by one's perceived relative deprivation, cognitive reappraisal and/or expressive suppression and empathy may play a chain-mediating role in relative deprivation and adolescent prosocial behavior. Therefore, the present study proposes hypotheses H5 and H6: Cognitive reappraisal and empathy have a chain-mediating effect on the association between relative deprivation and prosocial behavior (H5); expressive suppression and empathy have a chain-mediating effect on the association between relative deprivation and prosocial behavior (H6).

2.5. Summary

The association between relative deprivation and prosocial behavior in adolescents has not been examined, and the mechanisms underlying this association are unclear. By combining previous theoretical perspectives and empirical evidence, the present study sought to construct two chain mediation models. As shown in Figure 1, the objectives of the study were as follows: (a) to determine whether relative deprivation was a negative predictor of adolescent prosocial behavior; (b) to determine whether emotion regulation strategies and empathy act as chain mediators, with relative deprivation predicting cognitive reappraisal/expressive suppression, which in turn predicts empathy, and ultimately predicts adolescent prosocial behavior.

3. Methods

3.1. Participants

This study was conducted with high school students in a middle school in Fujian Province, China. Regarding the sample size required for structural equation modeling (SEM), Comrey and Lee (1992) stated that a sample size of 50 is very poor, 100 is poor, 200 is fair, 300 is good, and 500 is very good. The total number of questionnaire items in this study was 51. Therefore, based on the rule of thumb and the "10-times rule" (Hair et al., 2011), the sample size of this study should be greater than 510. Considering the risk of non-return of questionnaire, 650 questionnaires were distributed, and 633 questionnaires were successfully returned, with a return rate of 97.4%. Based on the regularity of questionnaire responses, invalid questionnaires (e.g., the same answer for nearly every question) were excluded, and 609 questionnaires were deemed valid, with an efficiency rate of 93.7%. In the valid samples, the average age was 15.42 years (SD = 0.653). In terms of gender distribution, 47.3% of participants were males and 52.7% were females, which closely resembled the gender ratio of high school students.

3.2. Instruments

3.2.1. Relative deprivation questionnaire

As previous studies have confirmed that cultural differences exist in relative deprivation (Smith et al., 2018), we used the Relative Deprivation Questionnaire (RDQ) developed by Ma (2012). The RDQ is based on the general population of China, and its items are aligned

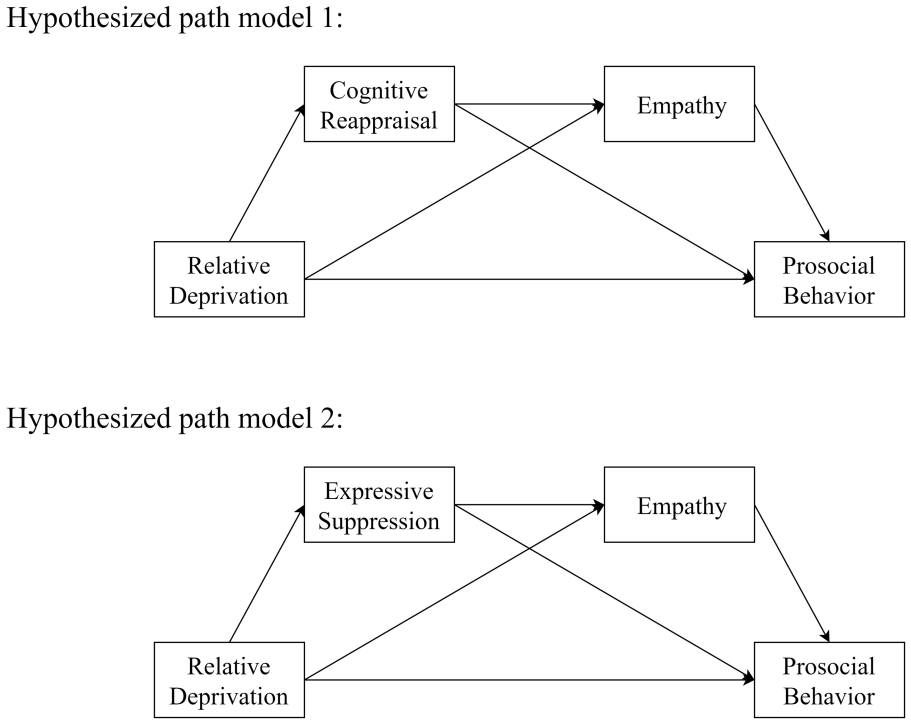


FIGURE 1
The hypothetical model diagram of this study.

with collectivist cultural perceptions of relative deprivation (Van den Bos et al., 2015), e.g., "Most of those rich people in society got rich through dishonorable means." It has been applied to adolescents in previous studies and demonstrated good psychometric properties (Yang et al., 2021). The RDQ consists of four items in a single dimension, scored using a 6-point scale that ranges from 1 (strongly disagree) to 6 (strongly agree). The higher the total score, the greater the relative deprivation. In the present study, the Cronbach's alpha coefficient for the RDQ was 0.704, indicating acceptable internal consistency (Taber, 2018).

3.2.2. Emotion regulation scale

We used the Emotion Regulation Scale (ERS), developed by Wang et al. (2007), to measure cognitive reappraisal and/or expressive suppression in adolescents. It is based on the emotion regulation model of Gross (1999) and has been revised to align with Chinese culture. The ERS has good reliability and measurement equivalence for the adolescent population (Chen et al., 2023), and it has been widely used to measure emotion regulation strategies in Chinese adolescents (Wang et al., 2022). The ERS consists of 14 items, including two dimensions (cognitive reappraisal/expressive suppression), with seven questions each. Participants rate their agreement with each statement using a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). Higher scores on each subscale indicate a greater tendency to use the corresponding emotion regulation strategy. In the present study, the Cronbach's alpha coefficients for the total scale, cognitive reappraisal subscale, and expressive suppression subscale were 0.833, 0.865, and 0.730, respectively, indicating good internal consistency.

3.2.3. Basic empathy scale

The Basic Empathy Scale (BES) was developed by Jolliffe and Farrington (2006) for adolescents. Compared to previous empathy measurement tools, the BES focused on both affect congruence (affective empathy) and understanding of another person's emotions (cognitive empathy) that more accurately conform to the concept of empathy (Jolliffe and Farrington, 2006). The present study used the Chinese version translated by Li et al. (2011) that has shown good psychometric properties in previous studies with Chinese adolescents (Yu et al., 2020). The BES consists of two dimensions and 20 items, specifically nine items on cognitive empathy and 11 items on affective empathy. Participants rate each item on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). Higher scores indicate higher levels empathy. In the present study, the Cronbach's alpha coefficients for the total scale, cognitive empathy subscale, and affective empathy subscale were 0.818, 0.789, and 0.757, respectively, indicating good internal consistency.

3.2.4. Prosocial behavior scales

Regarding the conceptual structure of the Chinese adolescents' Prosocial Behavior Scale (PBS), besides the altruism dimension reflecting purely the interest of others, it includes the compliance dimension reflecting adherence to societal norms or organizational rules, the relationship dimension reflecting interpersonal harmony in social interactions, and the personal trait dimension reflecting the motivation for self-improvement. Therefore, the PBS developed by Yang et al. (2016) can accurately measure Chinese adolescent prosocial behavior. It has been widely used to measure prosocial behavior in Chinese adolescents (Zhou et al., 2020). The PBS consists of altruism (four items), compliance (five items), relationship (three items), and

personal traits (three items). Participants rate each item on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). Higher scale scores indicate increased prosocial behavior. In the present study, the Cronbach's alpha coefficients for the total scale, altruism subscale, compliance subscale, relationship subscale, and personal trait subscale were 0.907, 0.754, 0.777, 0.631, and 0.724, indicating good internal consistency.

3.3. Procedures and statistical analysis

The present study used the cluster random sampling method. Students from 13 classes in the first and the second grades of high school were selected through a random number table. The test was administered by two graduate students majoring in psychology, and the instructions were read out after the questionnaires were distributed; the questionnaires were collected within the specified time.

After data collection, we calculated the descriptive statistics (i.e., mean score and standard deviation) and bivariate correlation of the variables using SPSS 25. The present study used self-report scales to collect the data that may lead to common method bias (Lindell and Whitney, 2001). All items were included in the exploratory factor analysis, according to Harman's single-factor test for the common method bias. Mplus 7.4 offers powerful analytical capabilities and flexibility in latent variable modeling, which can cope with multivariate, multidimensional, and multilevel analytic needs; it provides comprehensive model fitting and diagnostic information to help researchers gain a deeper understanding and interpretation of latent variable structures and relationships (Wang and Wang, 2019). Therefore, the following statistical analyses involving latent variables were conducted using Mplus 7.4. Referring to the guideline proposed by Rönkkö and Cho (2022), we used confirmatory factor analysis (CFA) to assess discriminant validity. As SEM has the advantages of independent variables containing measurement errors, high precision of parameter estimation, and rich evaluation indexes for model fitting, it was used to examine the mediating effect of cognitive reappraisal/expressive suppression and empathy on the association between relative deprivation and prosocial behavior. The PBS was parceled into four items according to the four subscales, and the BES was parceled into two items according to the two subscales. As suggested by Wu and Wen (2011), the single dimensional RDQ, the cognitive reappraisal subscale, and expressive suppression subscale were each parceled into two items using the odd-even method. The significance of the mediating effect was analyzed by the bootstrap method, with sampling for 5,000 times, and 95% confidence intervals were calculated; if the confidence interval did not include zero, it indicated a significant effect (Cheung, 2007).

4. Results

4.1. Common method bias and discriminant validity

For common method bias, the unrotated exploratory factor analysis extracted 11 factors with eigenvalues greater than 1. The first factor explained 17.87% of the variance (below the critical threshold of 40%), indicating that no serious common method biases were present in the data. For discriminant validity, as shown in Table 1, all correlations between the factors in the CFA were smaller than the square values of the average variance extracted (AVE) for each factor, thus satisfying the Fornell and Larcker (1981) criterion.

4.2. Correlation analysis between variables

Pearson zero-order correlations between the variables were calculated (Table 1). Relative deprivation was negatively correlated with cognitive reappraisal ($r = -0.143$, $p < 0.001$), empathy ($r = -0.149$, $p < 0.001$), and prosocial behavior ($r = -0.287$, $p < 0.001$), but positively correlated with expressive suppression ($r = 0.178$, $p < 0.001$). Cognitive reappraisal was positively correlated with expressive suppression ($r = 0.355$, $p < 0.001$), empathy ($r = 0.168$, $p < 0.001$), and prosocial behavior ($r = 0.279$, $p < 0.001$). Expressive suppression was not significantly correlated with empathy ($r = -0.075$, $p = 0.065$) and prosocial behavior ($r = -0.031$, $p = 0.447$). Empathy was positively correlated with prosocial behavior ($r = 0.442$, $p < 0.001$).

4.3. Chain-mediating effect of cognitive reappraisal and empathy

The hypothesized path model 1 comprised 10 observed variables and 4 latent variables (relative deprivation, cognitive reappraisal,

TABLE 1 Descriptive statistics and correlation coefficients for all variables.

Variables	Mean	SD	1	2	3	4	5
1 Relative Deprivation	12.50	4.06	0.762	−0.193***	0.235***	−0.295***	−0.365***
2 Cognitive Reappraisal	31.24	7.86	−0.143***	0.904	0.423***	0.283***	0.312***
3 Expressive Suppression	27.93	8.26	0.178***	0.355***	0.800	−0.009	−0.046
4 Empathy	70.10	9.63	−0.149***	0.168***	−0.075	0.632	0.607***
5 Prosocial Behavior	72.18	14.45	−0.287***	0.279***	−0.031	0.442***	0.858

Pearson zero-order correlations for all variables are presented below the diagonal, and factor correlations for all variables are presented above the diagonal. The diagonal is the square value of the Average Variance Extracted (AVE). According to the decision rule presented by Fornell and Larcker (1981), discriminant validity holds for two scales if the square values of the AVE for both are higher than the factor correlation between the scales.
***$p < 0.001$.

10.3389/fpsyg.2023.1238308

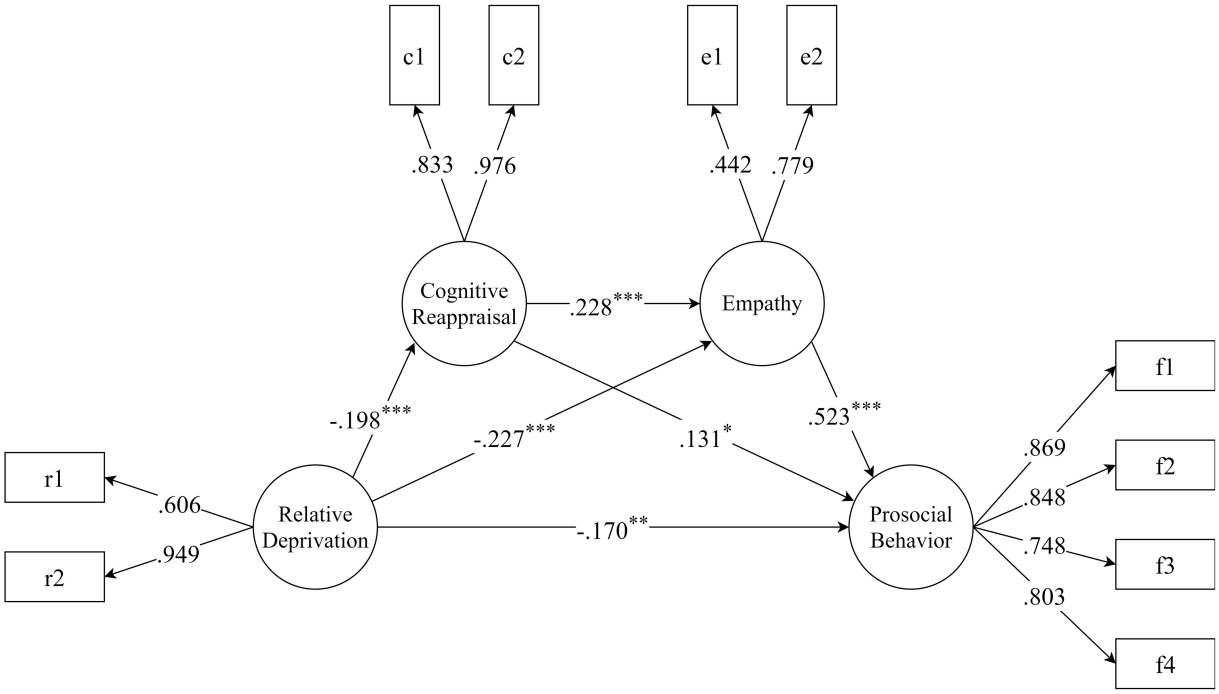


FIGURE 2
The path diagram of hypothesized model 1 (cognitive reappraisal and empathy as mediators). The structural equation model linking relative deprivation and prosocial behavior through cognitive reappraisal and empathy. The PBS was parceled into four items according to the four subscales, and the BES was parceled into two items according to the two subscales. As suggested by Wu and Wen (2011), the unidimensional RDQ and cognitive reappraisal subscale into two items using the odd-even method. Pathway coefficient and factors loadings are standardized. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 2 Bootstrapping estimates of 95% confidence intervals of (CI) estimation for the model pathways and indirect effect (model 1).

Model pathways	Boot lower 2.5%	Effect	Boot upper 2.5%	*p*-value
Relative deprivation → Prosocial behavior	−0.277	−0.170	−0.057	< 0.001
Relative deprivation → Cognitive reappraisal	−0.306	−0.198	−0.097	< 0.001
Relative deprivation → Empathy	−0.365	−0.227	−0.104	< 0.001
Cognitive reappraisal → Prosocial behavior	0.014	0.131	0.229	0.016
Cognitive reappraisal → Empathy	0.088	0.228	0.361	< 0.001
Empathy → Prosocial behavior	0.361	0.523	0.671	< 0.001
Mediating effect of cognitive reappraisal	−0.058	−0.026	−0.006	0.041
Mediating effect of empathy	−0.224	−0.119	−0.051	0.007
Chain-mediating effect of cognitive reappraisal and empathy	−0.055	−0.024	−0.009	0.021

empathy and prosocial behavior). This model showed an excellent fit with the data: $\chi^2/df = 3.902$, $RMSEA = 0.069$, $CFI = 0.969$, $TLI = 0.952$, and $SRMR = 0.031$. The results (Figure 2) showed that (1) relative deprivation negatively predicted cognitive reappraisal ($\beta = -0.198$, $p < 0.001$), empathy ($\beta = -0.227$, $p < 0.001$) and prosocial behavior ($\beta = -0.170$, $p = 0.003$); (2) cognitive reappraisal positively predicted empathy ($\beta = 0.228$, $p < 0.001$) and prosocial behavior ($\beta = 0.131$, $p = 0.016$); and (3) empathy positively predicted prosocial behavior ($\beta = 0.523$, $p < 0.001$). The bootstrapping estimates of 95% confidence intervals (CIs) of the model pathways and mediating effect are shown in Table 2. The mediating effect of cognitive reappraisal was significant (effect = −0.026, $p = 0.041$), accounting for 7.67% of the total effect. The mediating effect of empathy was significant (effect = −0.119, $p = 0.007$), accounting for 35.10% of the total effect. The chain-mediating effect of cognitive reappraisal and empathy was also significant (effect = −0.024, $p = 0.021$), accounting for 7.08% of the total effect.

4.4. Chain-mediating effect of expressive suppression and empathy

The hypothesized path model 2 comprised 10 observed variables and 4 latent variables (relative deprivation, expressive suppression,

10.3389/fpsyg.2023.1238308

empathy and prosocial behavior). This model showed an acceptable fit with the data: $\chi^2/df = 5.384$, $RMSEA = 0.085$, $CFI = 0.946$, $TLI = 0.916$, and $SRMR = 0.044$. The results (Figure 3) showed that (1) relative deprivation positively predicted expressive suppression ($\beta = 0.237$, $p = 0.002$), but negatively predicted empathy ($\beta = -0.308$, $p < 0.001$) and prosocial behavior ($\beta = -0.202$, $p = 0.002$); (2) expressive suppression not significantly predicted empathy ($\beta = 0.058$, $p = 0.535$) and prosocial behavior ($\beta = 0.005$, $p = 0.939$); (3) empathy positively predicted prosocial behavior ($\beta = 0.552$, $p < 0.001$). The bootstrapping estimates of 95% CIs of the model pathways and mediating effect are shown in Table 3. The mediating effect of expressive suppression was not significant (effect = 0.001, $p = 0.942$). The mediating effect of empathy was significant (effect = −0.170, $p = 0.001$), accounting for 46.68% of the total effect. The chain-mediating effect of expressive suppression and empathy was not significant (effect = 0.008, $p = 0.581$).

5. Discussion

The present study revealed that both cognitive reappraisal and empathy separately mediate the association between relative deprivation and adolescent prosocial behavior. Additionally,

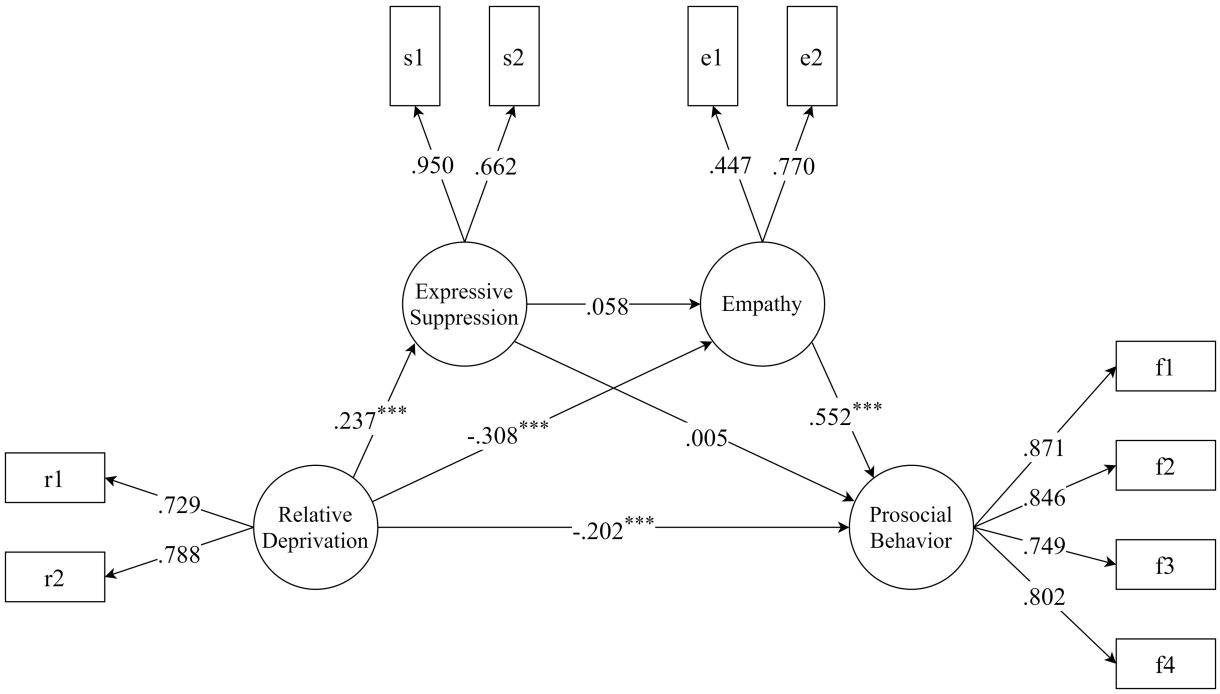


FIGURE 3
The path diagram of hypothesized model 2 (expressive suppression and empathy as mediators). The structural equation model linking relative deprivation and prosocial behavior through expressive suppression and empathy. The PBS was parceled into four items according to the four subscales, and the BES was parceled into two items according to the two subscales. As suggested by Wu and Wen (2011), the unidimensional RDQ and expressive suppression subscale were each parceled into two items using the odd-even method. Pathway coefficient and factors loadings are standardized. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 3 Bootstrapping estimates of 95% confidence intervals of (CI) estimation for the model pathways and indirect effect (model 2).

Model pathways	Boot lower 2.5%	Effect	Boot upper 2.5%	*p*-value
Relative deprivation → Prosocial behavior	−0.320	−0.202	−0.068	0.002
Relative deprivation → Expressive suppression	0.064	0.237	0.368	0.002
Relative deprivation → Empathy	−0.453	−0.308	−0.165	< 0.001
Expressive suppression → Prosocial behavior	−0.105	0.005	0.135	0.939
Expressive suppression→ Empathy	−0.149	0.058	0.213	0.535
Empathy → Prosocial behavior	0.361	0.552	0.673	< 0.001
Mediating effect of expressive suppression	−0.026	0.001	0.036	0.942
Mediating effect of empathy	−0.299	−0.170	−0.090	0.001
Chain-mediating effect of expressive suppression and empathy	−0.015	0.008	0.039	0.581

cognitive reappraisal and empathy chain-mediate the association between relative deprivation and adolescent prosocial behavior. However, expressive suppression was not found to mediate the association between relative deprivation and adolescent prosocial behavior.

5.1. Cognitive reappraisal rather than expressive suppression plays a mediating role

Supporting the hypothesis of this study, we found that relative deprivation significantly predicted cognitive reappraisal negatively, which is consistent with the results of Liu et al. (2021). This finding suggests that individuals with higher relative deprivation are less likely to use cognitive reappraisal when emotions need to be regulated. From the perspective of social comparison theory (Suls and Wheeler, 2012), when adolescents engage in comparisons with their peers, if they perceive themselves to be at a relative disadvantage or facing unfair circumstances, it activates negative thinking patterns and increases the likelihood of adopting negative coping strategies (Xiong et al., 2022).

The results showed that cognitive reappraisal significantly predicted prosocial behavior positively and mediated the association between relative deprivation and prosocial behavior, which is consistent with the hypothesis of this study. Cognitive reappraisal is considered an adaptive strategy that helps individuals perceive situations in a positive light and reduce the impact of negative emotions (Sun et al., 2020). According to the broaden-and-build theory of positive emotions (Fredrickson, 1998), positive emotional states play a role in expanding attention and awareness as well as promoting prosocial behavior. Previous empirical research has also confirmed the link between cognitive reappraisal and prosocial behavior (Li et al., 2021; Hodge et al., 2023). Thus, adolescents with high relative deprivation tend to use cognitive reappraisal strategies less frequently and, resultantly, are less likely to engage in prosocial behavior.

Interestingly, our results revealed that relative deprivation significantly predicted expressive suppression positively, but expressive suppression did not predict prosocial behavior. Moreover, there was no significant mediating effect of expressive suppression on the association between relative deprivation and prosocial behavior, a finding that contradicted the hypothesis of this study. According to the theory of emotion suppression proposed by Langner et al. (2012), when individuals with low social status subjectively believe that they are in a worse position than others, they usually choose to hide and suppress their negative emotions and behavioral expressions (i.e., using the expression suppression strategy) in order to prevent showing their dissatisfaction and anger in front of other individuals with higher status and reduce the risk of conflict between the disadvantaged and dominant groups. This theory explains the relationship between relative deprivation and expressive suppression. However, expressive suppression strategies do not mitigate negative emotional experience and only serve to inhibit the outward manifestations of negative behaviors (Gross and Cassidy, 2019). Thus, the use of expressive suppression strategies may not explain the effect of relative deprivation on prosocial behavior.

5.2. Chain mediation role of cognitive reappraisal and empathy

Consistent with the hypothesis of this study, we found that empathy significantly predicts prosocial behavior positively and mediated the association between relative deprivation and prosocial behavior. When adolescents experience relative deprivation, their self-concept is threatened (Kural and Kovács, 2022), which inhibits their empathic concern for others in need (Krol and Bartz, 2022). This lack of empathy may then lead to reduced prosocial behavior. In addition, individuals with high relative deprivation may also experience high levels of counter-empathy, such as envy and schadenfreude (Neufeld and Johnson, 2016), which could further inhibit empathy and decrease motivation for prosocial behavior.

Furthermore, the results showed that cognitive reappraisal and empathy chain-mediated the association between relative deprivation and prosocial behavior significantly. According to the integrative account of empathy and emotion regulation (Thompson et al., 2019), the way in which we understand and respond to others' emotions may be influenced by emotion regulation. Positive and adaptive emotion regulation strategies can reduce negative emotions during the empathy process and facilitate prosocial behavior (Lockwood et al., 2014). Cognitive reappraisal has been found to promote prosocial behavior by mediating empathy in previous studies (Laghi et al., 2018). In the empathy accuracy task, the use of cognitive reappraisal strategies improved individuals' accuracy in empathizing with others' negative emotions (Guo et al., 2023). Therefore, adolescents with high relative deprivation may struggle to empathize with others due to a lack of adaptive emotion regulation strategies that decreases the motivation to engage in prosocial behavior. It is worth noting that the chain-mediating effect was partially mediated the association between relative deprivation and prosocial behavior and accounted for a relatively small proportion of the total effect, suggesting that relative deprivation may also influence prosocial behavior through other mechanisms. A more comprehensive understanding of how relative deprivation affects prosocial behavior will be explored in future research.

Inconsistent with the hypothesis of this study, our results indicated that expressive suppression and empathy do not chain-mediate the association between relative deprivation and prosocial behavior. Expressive suppression did not significantly predict empathy and prosocial behavior, which may be influenced by contextual or cultural factors. Some studies have suggested that expressive suppression can be an effective emotion regulation strategy in certain contexts (Guo et al., 2023). For example, expressive suppression has been found to reduce negative emotion arousal more quickly than cognitive reappraisal among Chinese participants, although it also requires more cognitive resources (Yuan et al., 2015). Therefore, it is unclear whether the use of expressive suppression strategies can effectively regulate personal distress in response to negative stimuli during the empathy process. Currently, empirical findings on the effects of expressive suppression strategies are inconsistent, indicating the need for further exploration of potential moderators in future research.

5.3. Research implications

Theoretically, the present study provides evidence on the association between relative deprivation and prosocial behavior in the

adolescent populations. We explored the chain-mediating effects of two emotion regulation strategies on empathy and tested new mediating pathways that could reveal the intrinsic association between relative deprivation and prosocial behavior in adolescents. The results support the social comparison theory, the broaden-and-build theory of positive emotions, and integrative account of empathy and emotion regulation, providing new insights into the dynamics of prosocial behavior.

Practically, by understanding the effect of emotion regulation strategies and empathy on the association between relative deprivation and adolescent prosocial behavior, school psychologists can design intervention programs targeting the moral affect and prosocial behavior among secondary students. For example, a cognitive reappraisal mental health course could be implemented as an intervention for promoting prosocial behavior among adolescents, especially among those with high relative deprivation. In addition, cognitive reappraisal training can be provided as an adjunct to empathy training for adolescents.

5.4. Limitations

It should be noted that the present study has the following limitations. First, the results of this study were based on cross-sectional data, thus preventing the establishment of causal relationships between the variables. Therefore, future research should consider employing longitudinal designs for testing the hypothetical models. Second, the data pertaining to the variables were collected using subjective reporting methods. Therefore, objective data must be obtained by combining multiple methods such as parent, teacher, and peer evaluations to reduce the social praise effect. Additionally, the participants in the same school inevitably had consistent group characteristics, which may have affected the stability and generalizability of the present results. Whether the models hypothesized in this study holds true for a wider range of adolescents remains to be further tested. Finally, other variables that may influence the findings, such as the socioeconomic status of the study participants, were not collected in this study, and this aspect should be addressed in future studies.

6. Conclusion

In summary, the present study tested two chain-mediation models to explore the association between relative deprivation and adolescent prosocial behavior. Both cognitive reappraisal and empathy separately mediated the association between relative deprivation and adolescent prosocial behavior, whereas expressive suppression did not. Additionally, cognitive reappraisal and empathy chain-mediated the association between relative deprivation and adolescent prosocial behavior. Despite some limitations, the present study contributed to a better understanding of the association between relative deprivation and prosocial behavior, and the results emphasize the integrative role of empathy and emotion regulation as the underlying mechanism. This study also provides evidence that can form the basis for interventions that promote prosocial behavior among adolescents.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/bn9pm/files/osfstorage.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the School of Psychology, Fujian Normal University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

YX wrote the initial draft of this manuscript. SC and XS collected data for the study. DY supervised the project and provided critical revisions. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the National Education Science "Thirteenth Five-Year Plan" Key Project of the Ministry of Education (No. DBA190307), Education Reform Project of the Psychology Teaching Steering Committee in Higher Education Institutions under the Ministry of Education (No. 20222012), and the University-Industry Cooperation and Collaborative Education Project of the Ministry of Education (No. 220604497155844).

Acknowledgments

The authors would like to thank the editor and reviewers for their help in improving the quality of the article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Batson, C. D. (2017). "The empathy altruism hypothesis what and so what?" in *The Oxford handbook of compassion science*. eds. E. M. Seppälä, S. Simon-Thomas, S. L. Brown, M. C. Worline, C. D. Cameron and J. R. Doty (Oxford: Oxford University Press), 27–40.

Bebko, G. M., Franconeri, S. L., Ochsner, K. N., and Chiao, J. Y. (2011). Look before you regulate: differential perceptual strategies underlying expressive suppression and cognitive reappraisal. *Emotion* 11, 732–742. doi: 10.1037/a0024009

Benita, M., Levkovitz, T., and Roth, G. (2017). Integrative emotion regulation predicts adolescents' prosocial behavior through the mediation of empathy. *Learn. Instr.* 50, 14–20. doi: 10.1016/j.learninstruc.2016.11.004

Berman, J. Z., and Silver, I. (2022). Prosocial behavior and reputation: when does doing good lead to looking good? *Curr. Opin. Psychol.* 43, 102–107. doi: 10.1016/j.copsyc.2021.06.021

Callan, M. J., Kim, H., Gheorghiu, A. I., and Matthews, W. J. (2017). The interrelations between social class, personal relative deprivation, and prosociality. *Soc. Psychol. Personal. Sci.* 8, 660–669. doi: 10.1177/1948550616673877

Carlo, G., and Pierotti, S. L. (2020). "The development of prosocial motives" in *The Oxford handbook of moral development: an interdisciplinary perspective*. ed. L. A. Jensen (Oxford: Oxford University Press)

Chen, W., Gao, R., and Wang, L. (2023). Measurement invariance of emotion regulation scale in adolescents on gender, left-behind and time variables. *Chin. J. Clin. Psych.* 31, 112–115.

Cheung, M. W. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Struct. Equ. Model. Multidiscip. J.* 14, 227–246. doi: 10.1080/10705510709336745

Compas, B. E., Jaser, S. S., Bettis, A. H., Watson, K. H., Gruhn, M. A., Dunbar, J. P., et al. (2017). Coping, emotion regulation, and psychopathology in childhood and adolescence: a meta-analysis and narrative review. *Psychol. Bull.* 143, 939–991. doi: 10.1037/bul0000110

Comrey, A. L., and Lee, H. B. (1992). "Interpretation and application of factor analytic results" in *A first course in factor analysis*. eds. A. L. Comrey and H. B. Lee (Hillsdale: Lawrence Eribaum Associates)

Davis, M. H. (2015). "Empathy and prosocial behavior" in *The Oxford handbook of prosocial behavior*. eds. D. A. Schroeder and W. G. Graziano (Oxford: Oxford University Press)

Decety, J., and Michalska, K. J. (2010). Neurodevelopmental changes in the circuits underlying empathy and sympathy from childhood to adulthood. *Dev. Sci.* 13, 886–899. doi: 10.1111/j.1467-7687.2009.00940.x

El Mallah, S. (2020). Conceptualization and measurement of adolescent prosocial behavior: looking back and moving forward. *J. Res. Adolesc.* 30, 15–38. doi: 10.1111/jora.12476

Feather, N. T. (1999). *Values, achievement, and justice: Studies in the psychology of deservingness*. New York: Kluwer Academic/Plenum.

Feather, N. T., and Nairn, K. (2005). Resentment, envy, schadenfreude, and sympathy: effects of own and other's deserved or undeserved status. *Aust. J. Psychol.* 57, 87–102. doi: 10.1080/00049530500048672

Fornell, C., and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* 24, 337–346.

Fredrickson, B. L. (1998). What good are positive emotions? *Rev. Gen. Psychol.* 2, 300–319. doi: 10.1037/1089-2680.2.3.300

Fu, D., Qi, Y., Wu, H., and Liu, X. (2017). Integrative neurocognitive mechanism of empathy and counter-empathy. *Chin. Sci. Bull.* 62, 2500–2508. doi: 10.1360/N972016-01108

Gheorghiu, A. I., Callan, M. J., and Skylark, W. J. (2021). Having less, giving less: the effects of unfavorable social comparisons of affluence on people's willingness to act for the benefit of others. *J. Appl. Soc. Psychol.* 51, 946–961. doi: 10.1111/jasp.12813

Gross, J. J. (1999). Emotion regulation: Past, present, future. *Cognit. Emot.* 13, 551–573. doi: 10.1080/026999399379186

Gross, J. T., and Cassidy, J. (2019). Expressive suppression of negative emotions in children and adolescents: theory, data, and a guide for future research. *Dev. Psychol.* 55, 1938–1950. doi: 10.1037/dev0000722

Guo, X., Zheng, H., Ruan, D., Hu, D., Wang, Y., Wang, Y., et al. (2023). Associations between empathy and negative affect: effect of emotion regulation. *Acta Psychol. Sin.* 55, 892–904. doi: 10.3724/SP.J.1041.2023.00892

Hair, J. F., Ringle, C. M., and Sarstedt, M. (2011). PLS-SEM: indeed a silver bullet. *J. Mark. Theory Pract.* 19, 139–152. doi: 10.2753/MTP1069-6679190202

Hart, D., and Carlo, G. (2005). Moral development in adolescence. *J. Res. Adolesc.* 15, 223–233. doi: 10.1111/j.1532-7795.2005.00094.x

Hermann, A., Bieber, A., Keck, T., Vaitl, D., and Stark, R. (2014). Brain structural basis of cognitive reappraisal and expressive suppression. *Soc. Cogn. Affect. Neurosci.* 9, 1435–1442. doi: 10.1093/scan/nst130

Hodge, R. T., Guyer, A. E., Carlo, G., and Hastings, P. D. (2023). Cognitive reappraisal and need to belong predict prosociality in Mexican-origin adolescents. *Soc. Dev.* 32, 633–650. doi: 10.1111/sode.12651

Jolliffe, D., and Farrington, D. P. (2006). Development and validation of the basic empathy scale. *J. Adolesc.* 29, 589–611. doi: 10.1016/j.adolescence.2005.08.010

Klein, N. (2017). Prosocial behavior increases perceptions of meaning in life. *J. Posit. Psychol.* 12, 354–361. doi: 10.1080/17439760.2016.1209541

Krol, S. A., and Bartz, J. A. (2022). The self and empathy: lacking a clear and stable sense of self undermines empathy and helping behavior. *Emotion* 22, 1554–1571. doi: 10.1037/emo0000943

Kural, A. I., and Kovács, M. (2022). The association between attachment orientations and empathy: the mediation effect of self–concept clarity. *Acta Psychol.* 229:103695. doi: 10.1016/j.actpsy.2022.103695

Laghi, F., Lonigro, A., Pallini, S., and Baiocco, R. (2018). Emotion regulation and empathy: which relation with social conduct? *J. Genet. Psychol.* 179, 62–70. doi: 10.1080/00221325.2018.1424705

Lange, J., and Crusius, J. (2015). Dispositional envy revisited: unraveling the motivational dynamics of benign and malicious envy. *Personal. Soc. Psychol. Bull.* 41, 284–294. doi: 10.1177/0146167214564959

Langner, C. A., Epel, E. S., Matthews, K. A., Moskowitz, J. T., and Adler, N. E. (2012). Social hierarchy and depression: the role of emotion suppression. *J. Psychol.* 146, 417–436. doi: 10.1080/00223980.2011.652234

Leach, C. W., and Spears, R. (2009). Dejection at in-group defeat and schadenfreude toward second-and third-party out-groups. *Emotion* 9, 659–665. doi: 10.1037/a0016815

Li, C., Lv, R., Liu, J., and Zhong, J. (2011). The adaptation of basic empathy scale among Chinese adolescents. *Chin. J. Clin. Psych.* 19, 163–166.

Li, J., Yao, M., and Liu, H. (2021). From social support to adolescents' subjective well-being: the mediating role of emotion regulation and prosocial behavior and gender difference. *Child Indic. Res.* 14, 77–93. doi: 10.1007/s12187-020-09755-3

Lindell, M. K., and Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *J. Appl. Psychol.* 86, 114–121. doi: 10.1037/0021-9010.86.1.114

Liu, S., Chen, M., Zhang, N., Li, Y., and Xu, Z. (2021). Effect of relative deprivation on depression in college students: the chain mediating effect analysis. *Chin. J. Ergon.* 27, 22–27.

Lockwood, P. L., Seara-Cardoso, A., and Viding, E. (2014). Emotion regulation moderates the association between empathy and prosocial behavior. *PLoS One* 9:e96555. doi: 10.1371/journal.pone.0096555

Ma, A. (2012). Relative deprivation and social adaption: the role of mediator and moderator. *Acta Psychol. Sin.* 44, 377–387. doi: 10.3724/SP.J.1041.2012.00377

Nadler, J., Day, M. V., Beshai, S., and Mishra, S. (2020). The relative deprivation trap: how feeling deprived relates to symptoms of generalized anxiety disorder. *J. Soc. Clin. Psychol.* 39, 897–922. doi: 10.1521/jscp.2020.39.10.897

Nelson, D. A., and Crick, N. R. (1999). Rose-colored glasses: examining the social information-processing of prosocial young adolescents. *J. Early Adolesc.* 19, 17–38. doi: 10.1177/0272431699019001002

Neufeld, D. C., and Johnson, E. A. (2016). Burning with envy? Dispositional and situational influences on envy in grandiose and vulnerable narcissism. *J. Pers.* 84, 685–696. doi: 10.1111/jopy.12192

Orben, A., Tomova, L., and Blakemore, S. J. (2020). The effects of social deprivation on adolescent development and mental health. *Lancet Child Adolesc. Health* 4, 634–640. doi: 10.1016/S2352-4642(20)30186-3

Ornaghi, V., Conte, E., and Grazzani, I. (2020). Empathy in toddlers: the role of emotion regulation, language ability, and maternal emotion socialization style. *Front. Psychol.* 11:586862. doi: 10.3389/fpsyg.2020.586862

Pak, T. Y., and Babiarz, P. (2023). Relative deprivation and prosocial behavior: evidence from South Korea. *Soc. Sci. J.* 1–22. doi: 10.1080/03623319.2022.2151794

Pfattheicher, S., Nielsen, Y. A., and Thielmann, I. (2022). Prosocial behavior and altruism: a review of concepts and definitions. *Curr. Opin. Psychol.* 44, 124–129. doi: 10.1016/j.copsyc.2021.08.021

Power, S. A., Madsen, T., and Morton, T. A. (2020). Relative deprivation and revolt: current and future directions. *Curr. Opin. Psychol.* 35, 119–124. doi: 10.1016/j.copsyc.2020.06.010

Rönkkö, M., and Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organ. Res. Methods* 25, 6–14. doi: 10.1177/1094428120968614

Schäfer, J. Ö., Naumann, E., Holmes, E. A., Tuschen-Caffier, B., and Samson, A. C. (2017). Emotion regulation strategies in depressive and anxiety symptoms in youth: a meta-analytic review. *J. Youth Adolesc.* 46, 261–276. doi: 10.1007/s10964-016-0585-0

Shaver, P. R., Mikulincer, M., and Chun, D. S. (2008). "Adult attachment theory, emotion regulation, and prosocial behavior," in *Regulating Emotions: Culture, Social Necessity, and Biological Inheritance*, eds. M. Vandekerckhove, C. von Scheve, S. Ismer, S. Jung and S. Kronast (Malden, MA: Blackwell Publishing), 121–145.

Smith, H. J., Pettigrew, T. F., Pippin, G. M., and Bialosiewicz, S. (2012). Relative deprivation: a theoretical and meta-analytic review. *Personal. Soc. Psychol. Rev.* 16, 203–232. doi: 10.1177/1088868311430825

Smith, H. J., Ryan, D. A., Jaurique, A., Pettigrew, T. F., Jetten, J., Ariyanto, A., et al. (2018). Cultural values moderate the impact of relative deprivation. *J. Cross-Cult. Psychol.* 49, 1183–1218. doi: 10.1177/0022022118784213

Suls, J., and Wheeler, L. (2012). Social comparison theory. In LangeP. A. M. Van, A. W. Kruglanski and E. T. Higgins (Eds.), *Handbook of theories of social psychology* London: Sage Publications Ltd.

Sun, Y., Bo, S., and Lv, J. (2020). Brain network analysis of cognitive reappraisal and expressive suppression strategies: evidence from EEG and ERP. *Acta Psychol. Sin.* 52, 12–25. doi: 10.3724/SP.J.1041.2020.00012

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* 48, 1273–1296. doi: 10.1007/s11165-016-9602-2

Tang, N. (2015). The effects of empathy training on prosocial behavior of primary school students. (unpublished master's thesis). Hunan Normal University, Changsha, China.

Thompson, N. M., Uusberg, A., Gross, J. J., and Chakrabarti, B. (2019). Empathy and emotion regulation: an integrative account. *Prog. Brain Res.* 247, 273–304. doi: 10.1016/bs.pbr.2019.03.024

Van den Bos, K., Van Veldhuizen, T. S., and Au, A. K. (2015). Counter cross-cultural priming and relative deprivation: the role of individualism–collectivism. *Soc. Justice Res* 28, 52–75. doi: 10.1007/s11211-014-0230-6

Wang, L., Liu, H., Li, Z., and Du, W. (2007). Reliability and validity of emotion regulation questionnaire Chinese revised version. *Chin. J. Health Psychol.* 15, 503–505.

Wang, J., and Wang, X. (2019). *Structural equation modeling: applications using Mplus*. Hoboken: John Wiley & Sons.

Wang, W., and Wu, X. (2020). Mediating roles of gratitude, social support and posttraumatic growth in the relation between empathy and prosocial behavior among adolescents after the Ya'an earthquake. *Acta Psychol. Sin.* 52, 307–316. doi: 10.3724/SP.J.1041.2020.00307

Wang, D., Yuan, B., Han, H., and Wang, C. (2022). Validity and reliability of emotion regulation questionnaire (ERQ) in Chinese rural-to-urban migrant adolescents and young adults. *Curr. Psychol. J. Div. Perspect. Div. Psychol. Issues* 41, 2346–2353. doi: 10.1007/s12144-020-00754-9

Wu, Y., and Wen, Z. (2011). Item parceling strategies in structural equation modeling. *Adv. Psychol. Sci.* 19, 1859–1867.

Xiong, M., Hu, Z. Q., and Ye, Y. D. (2022). Association of relative deprivation with social withdrawal and its underlying mechanisms: a large cross-sectional study among Chinese migrant adolescents. *Curr. Psychol.* 42, 20849–20859. doi: 10.1007/s12144-022-03194-9

Yang, B., Cai, G., Xiong, C., and Huang, J. (2021). Relative deprivation and game addiction in left-behind children: a moderated mediation. *Front. Psychol.* 12:639051. doi: 10.3389/fpsyg.2021.639051

Yang, Y., Zhang, M., and Kou, Y. (2016). The revalidation and development of the prosocial behavior scale for adolescent. *Chin. Soc. Psychol. Rev.* 10, 135–150.

Yin, Y., and Wang, Y. (2023). Is empathy associated with more prosocial behaviour? A meta-analysis. *Asian J. Soc. Psychol.* 26, 3–22. doi: 10.1111/ajsp.12537

Yu, G., Li, S., and Zhao, F. (2020). Childhood maltreatment and prosocial behavior among Chinese adolescents: roles of empathy and gratitude. *Child Abuse Negl.* 101:104319. doi: 10.1016/j.chiabu.2019.104319

Yuan, J., Long, Q., and Ding, N. (2015). Suppression dampens unpleasant emotion faster than reappraisal: neural dynamics in a Chinese sample. *Sci. China Life Sci.* 58, 480–491. doi: 10.1007/s11427-014-4739-6

Zhang, N., Liu, W., Che, H., and Fan, X. (2023). Effortful control and depression in school-age children: the chain mediating role of emotion regulation ability and cognitive reappraisal strategy. *J. Affect. Disord.* 327, 111–119. doi: 10.1016/j.jad.2023.01.129

Zhang, H., Liu, M., and Tian, Y. (2016). Individual-based relative deprivation (IRD) decreases prosocial behavior. *Motiv. Emot.* 40, 655–666. doi: 10.1007/s11031-016-9564-8

Zhang, L., Qiao, L., Xu, M., Fan, L., Che, X., Diao, L., et al. (2021). Personal relative deprivation impairs ability to filter out threat-related distractors from visual working memory. *Int. J. Psychophysiol.* 162, 86–94. doi: 10.1016/j.ijpsycho.2021.02.008

Zhao, H., and Zhang, H. (2022). How personal relative deprivation influences moral disengagement: the role of malicious envy and honesty–humility. *Scand. J. Psychol.* 63, 246–255. doi: 10.1111/sjop.12791

Zhou, X., Hu, S., Liang, L., Yuan, K., and Bian, Y. (2020). Prosocial behavior and subjective well-being in junior high school students: a cross-lagged analysis during three years. *Chin. J. Clin. Psych.* 28, 561–565.

Zhou, S., Wu, Y., and Xu, X. (2023). Linking cognitive reappraisal and expressive suppression to mindfulness: a three-level Meta-analysis. *Int. J. Environ. Res. Public Health* 20:1241. doi: 10.3390/ijerph20021241

Zitek, E. M., Jordan, A. H., Monin, B., and Leach, F. R. (2010). Victim entitlement to behave selfishly. *J. Pers. Soc. Psychol.* 98, 245–255. doi: 10.1037/a0017168

frontiers | Frontiers in Psychology

TYPE Correction
PUBLISHED 06 August 2024
DOI 10.3389/fpsyg.2024.1466931

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
Delin Yu
yu.delin@foxmail.com

RECEIVED 18 July 2024
ACCEPTED 22 July 2024
PUBLISHED 06 August 2024

CITATION
Xu Y, Chen S, Su X and Yu D (2024) Corrigendum: Cognitive reappraisal and empathy chain-mediate the association between relative deprivation and prosocial behavior in adolescents.
Front. Psychol. 15:1466931.
doi: 10.3389/fpsyg.2024.1466931

Corrigendum: Cognitive reappraisal and empathy chain-mediate the association between relative deprivation and prosocial behavior in adolescents

Yanfeng Xu[1], Sishi Chen[1], Xiaojie Su[1,2] and Delin Yu[1]*

[1]School of Psychology, Fujian Normal University, Fuzhou, Fujian, China, [2]Normal College, Urumqi Vocational University, Urumqi, Xinjiang, China

KEYWORDS

relative deprivation, cognitive reappraisal, expressive suppression, empathy, prosocial behavior

A corrigendum on
Cognitive reappraisal and empathy chain-mediate the association between relative deprivation and prosocial behavior in adolescents

by Xu, Y., Chen, S., Su, X., and Yu, D. (2023). *Front. Psychol.* 14:1238308. doi: 10.3389/fpsyg.2023.1238308

In the published article, there was an error in the Funding statement. The grant number stated for the "Education Reform Project of Psychology Teaching Steering Committee of Higher Education Ministry of Education (No. 2022012)" was incorrect. The correct Funding statement appears below.

Funding

This work was funded by the National Education Science "Thirteenth Five-Year Plan" Key Project of the Ministry of Education (No. DBA190307), Education Reform Project of the Psychology Teaching Steering Committee in Higher Education Institutions under the Ministry of Education (No. 20222012), and the University-Industry Cooperation and Collaborative Education Project of the Ministry of Education (No. 220604497155844).

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's note

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 11 December 2023
DOI 10.3389/fpsyg.2023.1251276

OPEN ACCESS

EDITED BY
Paola Magnano,
Kore University of Enna, Italy

REVIEWED BY
Gabriele Chierchia,
University of Pavia, Italy
Hillary Schaefer,
Lynch Research Associates, United States

*CORRESPONDENCE
Chunhui Qi
✉ qchizz@126.com

RECEIVED 01 July 2023
ACCEPTED 27 November 2023
PUBLISHED 11 December 2023

CITATION
Zhang Z, Li M, Liu Q, Chen C and Qi C (2023) Group membership and adolescents' third-party punishment: a moderated chain mediation model.
Front. Psychol. 14:1251276.
doi: 10.3389/fpsyg.2023.1251276

Group membership and adolescents' third-party punishment: a moderated chain mediation model

Zhen Zhang[1,2], Menghui Li[3], Qiyun Liu[1], Chao Chen[4] and Chunhui Qi[1]*

[1]Faculty of Education, Henan Normal University, Xinxiang, China, [2]Faculty of Education, Henan University, Kaifeng, China, [3]Mental Health Education Center, Nanyang Medical College, Nanyang, China, [4]Zhumadian Basic Teaching Research Office, Zhumadian, China

Third-party punishment (TPP) reflects people's social preference for fairness norms and is fundamental to maintaining fairness norms on a large scale. Several empirical studies have shown that the offender's group membership impacts TPP, but the detailed mechanisms have yet to be fully elucidated. The current study used the third-party punishment game task to explore the relationship between group membership, perceived unfairness, anger, and adolescents' TPP. A total of 306 teenagers aged 12 to 15 were chosen as subjects through cluster sampling. The results showed that group membership (classmate vs. stranger) and gender can affect adolescents' TPP together, which manifests as adolescents enacting significantly harsher punishments on strangers than on classmates, especially for boys. Group membership indirectly affects TPP through the mediating effects of perceived unfairness, anger and through a chain mediation of perceived unfairness and anger. Moreover, gender positively moderate the relationship between group membership and perceived unfairness. Specifically, group membership significantly affects boys' perceived unfairness, but cannot predict girls' perceived unfairness. The above results can be used to guide adolescents toward appropriate justice concepts and moral awareness, thus enhancing TPP.

KEYWORDS

group membership, third-party punishment, unfair perception, anger, adolescents

1 Introduction

As an important way to safeguard social fairness, third-party punishment (TPP) refers to behavior in which individuals voluntarily provide resources to punish violators in response to irregularities (Fehr and Gächter, 2002). Behavioral economists and evolutionary psychologists emphasize that TPP can effectively suppress potential non-cooperative behavior, which is not only beneficial to the establishment and maintenance of long-term relationships but also helps to promote and maintain stability and harmony in society (Buckholtz and Marois, 2012). Scholars often use the third-party punishment game (TPPG) to explore how individuals deal with violations that do not involve their own interests and the factors that impact them. During this task, unrelated third-party participants observed an individual (i.e., a transgressor) providing an unfair distribution to a recipient (i.e., give $2 out of $10 to the recipient and keep $8 for yourself), and then decided whether to punish the selfish transgressor at their expenses (Fehr and Fischbacher, 2004). People across diverse societies have a willingness to punish unfair

players (Henrich et al., 2006; House et al., 2020), and this behavior is crucial for maintaining social cooperation (Balliet et al., 2014; Henrich and Muthukrishna, 2021).

The importance of TPP has attracted attention in many disciplines due to its role in promoting group cooperation and maintaining social order (Krueger and Hoffman, 2016; Marshall and McAuliffe, 2022). To understand the origin and development of third-party' fairness consideration, several studies have examined the TPP of children at different developmental stages (Gummerum and Chu, 2014; McAuliffe et al., 2015; Gummerum et al., 2016, 2020, 2022; Lee and Warneken, 2022). Six-year-old children begin to exhibit costly TPP (McAuliffe et al., 2015; Riedl et al., 2015; Salali et al., 2015), and the punishment pattern fully develops until 13–14 years of age (Bašić et al., 2020) and has a certain cross-cultural stability (House et al., 2020). However, to date, most related studies have examined children and adults as third-party punishers, and few studies have examined adolescents (Gummerum et al., 2020, 2022). Adolescence, defined as the period from 10 to 24 years of age (Sawyer et al., 2018), is characterized by heightened affective and social sensitivity (Towner et al., 2023). Moreover, group influence is highly prevalent during adolescence, which made adolescence more concerned with conformity and fitting in with others (Blakemore, 2018). Accordingly, adolescents may show exhibit more intense TPP than children and adults. Ultimately, there is a need to explore the factors that influence TPP among juveniles.

1.1 Group membership and TPP

Group membership is a social dimension that distinguishes oneself from others, including friendship, race, class, nationality, and even mere membership triggered by artificial cues (Lieberman and Linke, 2007; Chierchia et al., 2020; Zhang et al., 2021, 2022). Researchers have examined the effect of group membership on TPP, but their findings have been inconsistent. Two competing hypotheses, the *Mere Preferences Hypothesis* and the *Norms Focused Hypothesis* (McAuliffe and Dunham, 2016; Zhang et al., 2020), were developed to explain the contradictory results. The *Mere Preferences Hypothesis* suggests that individuals' positive evaluation toward the ingroup would reduce TPP for ingroup perpetrators, supported by the majority of evidence based on adults (Yudkin et al., 2016; McAuliffe et al., 2017; Guo et al., 2020, 2022; Yang et al., 2023) and children (Jordan et al., 2014), which supports the ingroup favoritism phenomenon (IGF). The *Norms Focused Hypothesis* emphasizes that individuals' concern for maintaining norms of group cooperation would enhance TPP for ingroup violators, as demonstrated by some evidence based on adults (Mendoza et al., 2014; Delton and Krasnow, 2017) and children (Gonzalez-Gadea et al., 2022), known as the black sheep effect (BSE). Even though they differ in the direction of the effect, IGF and BSE are two ways for people to maintain the group identity and cohesion (Zhang et al., 2020). However, various systematic reviews and meta-analyses have found that children, adolescents, and adults are more likely to punish outgroup offenders than ingroup criminals (McAuliffe and Dunham, 2016; Lazić et al., 2021). Therefore, group membership can influence adolescents' TPP, showing that youth punish outgroup members more harshly than in-group members (Hypothesis 1).

1.2 Perceived unfairness as a potential mediator

Perceived unfairness is one potential explanation for the proposed effect of group membership on TPP (Lu and McKeown, 2018). Fairness preference theory suggests that humans have a strong disgust for inequality and are willing to consume resources to punish offenders when they experience or witness injustices (Fehr and Schmidt, 1999). Some studies have shown that adolescents have strong equity concerns and a high willingness to sacrifice their personal interests to uphold fairness norms; thus, perceived unfairness has become an important driving force for implementing punishment (Güth and Kocher, 2014; Lu and McKeown, 2018). Nevertheless, the perception of injustice is not invariable, and it will depend on the group relationship of both sides. Firstly, individuals' perception of injustice is less prominent when unfair proposals are made by ingroups than by outgroups (Lu and McKeown, 2018). In addition, perceived unfairness is associated with TPP, such that the greater the perceived unfairness, the more motivated people are to punish (Fehr and Gächter, 2002). Finally, self-reported justice perception mediates the relationship between partners' social distance (i.e., human vs. computer partner) and rejection behavior among healthy adults and patients with major depressive disorders (Wang and Li, 2013; Jin et al., 2022). Therefore, we proposed that group membership may be associated with more TPP for outgroup members via increased perceived unfairness (Hypothesis 2).

1.3 Anger as a potential mediator

According to negative emotion theory, perceiving negative emotions such as anger, frustration, and disgust that arise from behavior violations can form the basis for punishing behaviors, triggering a desire to punish others in response to real-life immorality (Hartsough et al., 2020). Self-reported anger has been suggested as a possible motivation for TPP in some studies (Fehr and Fischbacher, 2004; Gummerum et al., 2016) and could mediate the association between unfair offers and TPP in adults (Gummerum et al., 2020). Additionally, the harshness of TPP increased significantly when anger was induced but decreased when anger was inhibited (Gummerum et al., 2022). More importantly, the experience of anger caused by injustice differs depending on the peer group to which one belongs. Bicskei et al. (2016) found that the same unkind behavior by outgroups was associated with greater anger-like emotions than that of ingroups, and punishment behavior was strongly influenced by anger-like emotions. Finally, Wang and Li (2013) found that self-reported feelings of anger could mediate the association between social relations (i.e., friend, teacher, and stranger) and adults' punishment in the ultimatum game. Hence, we proposed that anger played a crucial role in the relationship between group membership and TPP (Hypothesis 3).

1.4 Perceived unfairness and anger

An evaluation-emotional-behavioral model was employed by Seip et al. (2014) to explain the mechanism underlying costly punishments of unfairness: evaluating an action or event as unjust leads to anger

toward the offender, which can then drive people to punish those who violate social norms, even if punishment comes at a price. Several studies have indicated that violations of fairness can lead to perceived unfairness, which leads to anger and, ultimately motivates punishment by second and third parties (Singer and Steinbeis, 2009; Mendoza et al., 2014). For example, Mendoza et al. (2014) demonstrated that increasingly unfair offers predicted lower perceived fairness, thereby resulting in a strengthened level of anger and ultimately prompting individuals to reject the unfair offers in ultimatum game. Accordingly, we proposed that perceived unfairness and anger can exert a chain-mediating effect between group membership and TPP (Hypothesis 4).

1.5 Gender as a potential moderator

Social role theory emphasizes that different societal stereotypes are assigned to boys and men as compared to girls and women, with girls expected to be more communal and caring and boys are expected to be agentic and dominant (Eagly, 2009). These gender role beliefs greatly influence boys' and girls' perceptions, emotional experiences, and behavioral responses to norm-violating behavior (Chawla et al., 2020). Laboratory and field studies suggest that boys are more likely to judge private behavior negatively, experience greater anger, and punish offenders more severely when they experience normal transgressions than girls (Kromer and Bahçekapili, 2010; Bonini et al., 2011; Balafoutas and Nikiforakis, 2012; Rodriguez-Ruiz et al., 2019). As a result of this socialization and other forces, boys and men tend to be more socially dominant than girls and women (Pratto et al., 1994; Du et al., 2021), and norms for masculinity are more rigid than norms for femininity (Koenig, 2018). Accordingly, boys show greater ingroup favoritism both cognitively, emotionally and behaviorally than girls during intergroup interactions. For example, boys exhibit greater ingroup favoritism than girls when responding to unfair distributions (Wu and Gao, 2018). Thus, we proposed that the chain mediation of perceived unfairness and anger was more pronounced in boys than in girls (Hypothesis 5).

2 Method

2.1 Participants and procedure

The current study adopted a complete between-subject design of 2 (group membership: classmate vs. stranger) × 2 (gender: boy, girl). Based on an *a priori* power analysis, the sample size was estimated using G*Power 3.1 (Faul et al., 2007). F tests and ANOVA (fixed effects, special effects, main effects, and interactions) in G*Power (version 3.1.9.7) were selected. To detect a medium effect ($f^2 = 0.25$), $N = 128$ participants (32 participants per group) with 0.80 power and 0.05 Types I error rates were needed. Experimental data were collected from two junior high schools in Henan Province, China. The distribution and collection of situational questionnaires were conducted by a trained research assistant with standardized processes for completing the questionnaires. During the study, eight classes of seventh- and eighth-graders were randomly selected. Four classes were randomly assigned to the classmate condition, and the other four classes were assigned to the stranger condition.

A total of 350 questionnaires were distributed in the form of class tests. After removing missing values or other ineffective responses, the final data set consisted of 306 questionnaires, with a minimum of 49 respondents for each condition. The sample included 175 boy students (57.19%) and 131 girl students (42.81%) between the ages of 12 and 15. Their average age was 13.46 ± 0.75 years, with 69.99% in seventh grade and 33.01% in eighth grade. All subjects self-reported no mental or psychological disorders and gave their oral informed consent. Ethics committee approval was obtained from the Faculty of Education at Henan Normal University, and protocol adherence to the Declaration of Helsinki was ensured.

2.2 Experimental procedure and materials

Students were instructed to complete a pen-and-paper test on TPP in the classroom as a class. There were four main sections of the assessment, including basic personal information, third-party punishment tasks, group membership manipulation and check, and self-report assessment. These four sections were always administered in the same order as below.

2.2.1 Third-party punishment game

Based on the third-party punishment game paradigm designed by Fehr and Fischbacher (2004), a situational questionnaire was developed and administered as follows:

> To celebrate the National Day of China, your school held a literary and artistic performance. Two students, Li Ming and Wang Hua, collaborated in singing "I and My Motherland" and won first place in the competition. The school awarded a cash award of 100 (RMB) to the winning team. Li Ming, as a representative, went on stage to receive the award. The judge teacher reminded Li Ming that Wang Hua also contributed to this award and asked the two of them to share the award, allowing Li Ming to decide on how to allocate the money. Li Ming then provided an allocation scheme of 80:20, which means that Li Ming received 80 RMB, while Wang Hua received 20 RMB.

2.2.2 Group membership manipulation and check

In accordance with a previous study (Guo et al., 2016), we manipulate group membership by asking participants to imagine that the offender (Li Ming) is a classmate of theirs (ingroup condition) vs. is from a different class (outgroup condition). In both cases, the third-party victim (Wang Hua) was depicted as a stranger, both to the participants and to the offender. The Inclusion of Other in the Self (IOS) scale developed by Aron et al. (1992) was used to assess the perceived social distance between the two parties, and then the effectiveness of group membership manipulation was tested. The scale mainly uses the size of the overlapping area of two circles to determine the degree of closeness between the two circles, ranging from a complete distance of 1 point to an approximate overlap of 5 points. This article uses the Likert 5-point scoring method; the higher the score, the higher the degree of social distance.

2.2.3 Self-report assessment

Participants were told that imagined themselves as bystanders in the above scenario and assessed the following three aspects: (1) unfairness perception, that is, the unfairness degree of the allocation scheme of 80:20, measured on a scale of 1 to 7, with higher scores

representing higher perceived unfairness (Lu and McKeown, 2018); (2) anger, that is, how angry you feel about the unfair allocation, measured on a scale of 1 to 7, with higher scores representing more anger (Gummerum et al., 2020); and (3) punishment intensity, that is, the amount of punishment the participants are willing to impose, measured on a scale from 0 to 4, with a punishment ratio of 1:20 (each punishment amount will reduce the offender by 20 yuan) (Chen and Bo, 2016).

3 Results

3.1 Manipulation check

A 2 (group membership: classmate vs. stranger) × 2 (gender: boy vs. girl) ANOVA on the IOS scale scores showed that only the main effect of group membership was significant, $F(1,302) = 588.52$, $p < 0.01$, partial $\eta^2 = 0.66$. Identification with a classmate was larger ($M = 3.50$, $SE = 0.07$) than identification with stranger ($M = 1.42$, $SE = 0.06$), see Figure 1A. This finding indicated that the manipulation of group membership was successful.

3.2 Preliminary analyses

A 2 (group membership: classmate vs. stranger) × 2 (gender: boy vs. girl) MANOVA on perceived unfairness, anger, and TPP found that the main effects of group membership were significant, $Fs(1,302) > 19.75$, $ps < 0.01$, partial $\eta^2s > 0.06$. As compared to strangers' selfish behavior, participants perceived unfair allocation from classmates as less unfair, experienced less anger, and punished less severely. Moreover, the interactions by group membership and gender were also significant, $Fs(1,302) > 6.13$, $ps < 0.05$, partial $\eta^2s > 0.02$. Further analysis shown that girls' TPP was influenced by group membership, $F(1,302) = 7.67$, $p < 0.01$, but not by their perceptions of unfairness and anger, $Fs(1,302) < 1.82$, $ps > 0.05$. In particular, girls punished classmates ($M = 2.08$, $SE = 0.12$) less severely than stranger ($M = 2.51$, $SE = 0.10$). In contrast, boys' perceptions of unfairness, anger, and TPP were affected by group membership, $Fs(1,302) > 29.23$, $ps < 0.01$, showing that boys perceive classmate' transgressions as less unfair, experience less anger, and impose softer punishments comparing to stranger's transgressions (see Figures 1B–D). The main effects of gender were not significant, $Fs(1,302) < 2.69$, $ps > 0.05$.

The descriptive statistics and correlations of the variables are reported in Table 1. Dummy codes were used for group membership, with ingroup coded as 0 and outgroup coded as 1. Group membership was significantly positively associated with perceived unfairness, anger, and TPP ($r = 0.28, 0.28, 0.38$, $ps < 0.01$), supporting Hypothesis 1. Unfair perception was significantly positively associated with anger and TPP ($r = 0.68, 0.67$, $ps < 0.01$). Anger was significantly positively correlated with TPP ($r = 0.67$, $p < 0.01$).

3.3 Moderated chain mediation model

A moderated chain mediation model was conducted by using Model 85 in the Process 4.0 macro of SPSS 26.0. Dummy codes were used for gender and group membership, with girl and ingroup coded as 0 while boy and outgroup coded as 1. Confounding effects were reduced by including age and grade as control variables. The results showed that group membership could significantly positively predict unfair perception, anger and TPP ($\beta = 0.27, 0.09, 0.18$, $ps < 0.01$); unfair perception could significantly positively predict anger and TPP ($\beta = 0.67, 0.36$, $ps < 0.01$); and anger could significantly positively predict TPP ($\beta = 0.37$, $p < 0.01$). Thus, Hypothesis 2, 3 and 4 were supported. The interaction between group membership and gender had a significant effect on perceived unfairness ($\beta = 0.14$, $p < 0.05$). In contrast, the interaction between group membership and gender had no effect on anger ($\beta = 0.04$, $p > 0.05$) and TPP ($\beta = 0.03$, $p > 0.05$). Thus, Hypothesis 5 was partially supported (see Table 2).

A slope test was conducted to clarify the mechanisms by which group membership and gender interact with perceived unfairness. The result showed that group membership significantly positively predict the boys' perceived unfairness (simple slope = 0.39, $t = 5.47$, $p < 0.01$), but cannot predict girls' perceived unfairness (simple slope = 0.11, $t = 1.23$, $p > 0.05$) (see Figure 2A). The figures of the chain mediation model separately for boys and girls were shown in Figures 2B,C.

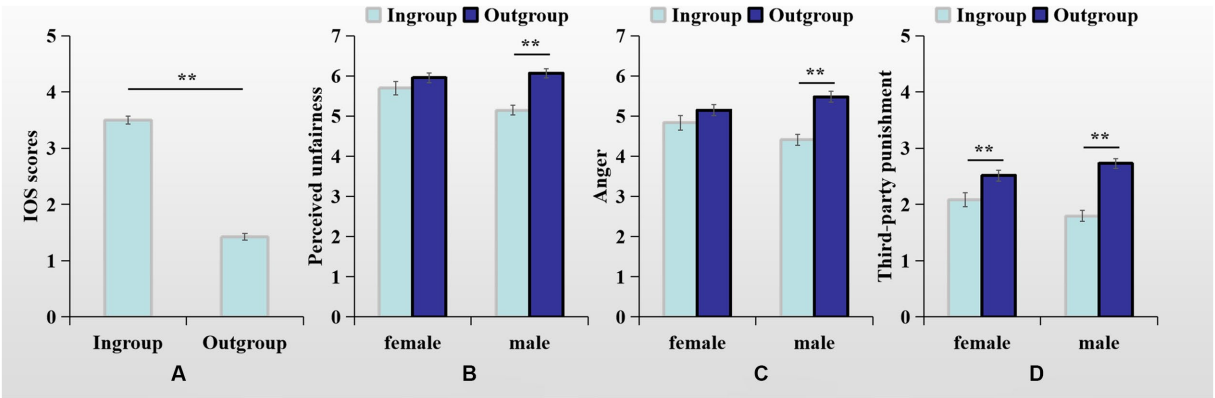


FIGURE 1
(A) IOS scores as a function of group membership; **(B)** Perceived unfairness, **(C)** anger and **(D)** third-party punishment as a function of group membership, separately for girls and boys. Error bars indicate standard error. *$p < 0.05$, **$p < 0.01$.

TABLE 1 Descriptive statistic and correlations of variables ($N = 306$).

Variable	*M*	SD	1	2	3	4	5
1. Gender	0.57	0.50	–	–	–	–	–
2. Group membership	0.57	0.50	−0.10	–	–	–	–
3. Perceived unfairness	5.73	1.18	−0.10	0.28**	–	–	–
4. Anger	5.00	1.33	−0.02	0.28**	0.68**	–	–
5.Third-party punishment	2.31	0.94	−0.04	0.38**	0.67**	0.67**	–

Gender and group membership is a virtual variable, female and ingroup = 0, male and outgroup = 1; *$p < 0.05$, **$p < 0.01$.

TABLE 2 The moderated chain mediating effect of perceived unfairness and anger.

Regression equation		Overall fitting index			Regression coefficient	
Result variable	Prediction variable	*R*	R^2	*F*	*β*	*t*
Perceived unfairness	Group membership	0.33	0.11	7.44**	0.27	4.90**
	Gender				−0.08	−1.48
	Group membership × Gender				0.14	2.56*
Anger	Perceived unfairness	0.71	0.50	50.08**	0.67	15.39**
	Group membership				0.09	2.17*
	Gender				0.05	1.28
	Group membership × Gender				0.04	1.03
Third-party punishment	Anger	0.75	0.56	54.08**	0.37	6.83**
	Perceived unfairness				0.36	6.54**
	Group membership				0.18	4.55**
	Gender				0.03	0.77
	Group membership × Gender				0.03	0.78

All variables in the model are brought back into the equation after standardized processing. Gender and group membership is a virtual variable, female and ingroup = 0, male and outgroup = 1; *$p < 0.05$, **$p < 0.01$.

4 Discussion

The current study explores the relationship between group membership and adolescent's TPP and its potential mechanism. The findings show that group membership and gender could affect adolescents' perceived unfairness, anger, and TPP together, which manifests as boys perceive classmates' transgressions as less unfair, experience less anger, and impose softer punishments compared to strangers' transgressions. Furthermore, group membership weakens adolescent's TPP through perceived unfairness, anger and a chain mediating path of perceived unfairness and anger, especially for boys.

Our results support the *Mere Preferences Hypothesis*, because adolescents' perceived unfairness, anger, and TPP both exhibited IFG instead of BSE. These findings are aligned with previous research based on adults (Yudkin et al., 2016; McAuliffe et al., 2017; Guo et al., 2020, 2022; Yang et al., 2023) and children (Jordan et al., 2014), which indicated that people are more likely to forgive ingroup offenders than outgroup offenders. From the perspective of psychological development, the replicated IFG effect in adolescents not only extends previous studies, but also coincides with recent meta-analysis results (Lazić et al., 2021). In other words, adolescents, like children and adults, care about and defend their group membership and are willing to forgive in-group violators. However, Gonzalez-Gadea et al. (2022) found that children aged 6 to 9 exhibited an ingroup policing bias but not an ingroup favoritism bias. One potential explanation for the difference is the cost of punishment. Yudkin et al. (2019) found that costly punishment, as a more effective way of group regulation, produces ingroup policing effects, rather than non-costly punishment. The TPP decision used in our study involves costless self-reported punishment, which might lead to IGF instead of BSE.

Moreover, perceived unfairness mediates the relationship between group membership and TPP for junior school students. In particular, outgroup infractions are perceived as more unjust in comparison to ingroup violations, thereby promoting TPP. Consistent with previous research (McCall et al., 2014; Lu and McKeown, 2018), the perception of injustice is comparatively less pronounced when inequitable propositions originate from ingroup as opposed to outgroup, regardless of whether the resource allocation scenario involves second or third parties. Based on the *Mere Preferences Hypothesis* (McAuliffe and Dunham, 2016), the identity of groups may lead to a positive appraisal and partiality toward ingroups, thereby fostering greater inclusivity toward ingroup offenders. Brain imaging research has suggested that individuals utilize mentalizing networks to comprehend and justify transgressions committed by ingroup members, which subsequently leads to weaker perceived unfairness (Baumgartner et al., 2012; Fatfouta et al., 2018). Furthermore, this aligns with prior studies that have demonstrated the role of perceived injustice as a mediator in the association between social distance and retribution enacted by a second party (Wang and Li, 2013; Jin et al., 2022). Thus, in comparison to classmates, third-party bystanders tend to view transgressions committed by strangers as more unjust, which subsequently results in severe TPP.

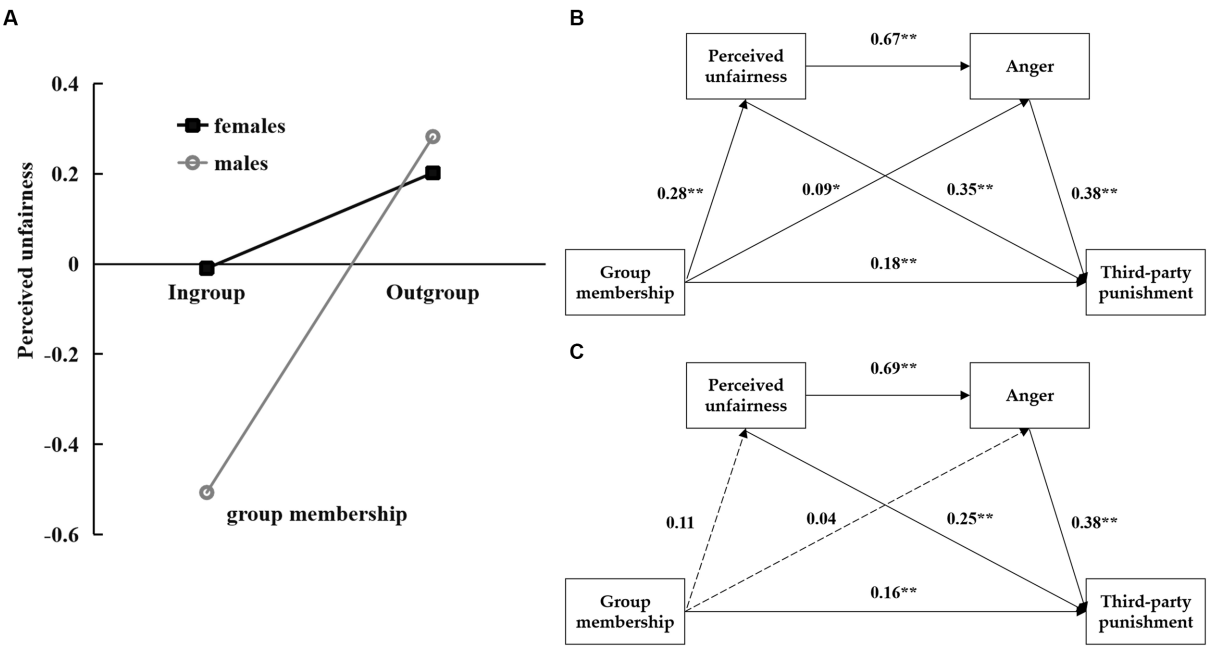


FIGURE 2
The moderating role of gender in the relation between group membership and perceived unfairness **(A)**; The figures of the chain mediation model separately for boys **(B)** and girls **(C)**.

Once more, the relationship between group membership and adolescents' TPP is mediated by anger. Specifically, strangers' infraction triggers stronger anger than classmates' infraction, leading to more severe punishment. This also supports the *Mere Preferences Hypothesis*, showing that ingroup violations are emotionally tolerated by people (McAuliffe and Dunham, 2016). As previously demonstrated (McCall et al., 2014; Bicskei et al., 2016), anger emotions were less salient when unfair allocations were provided by ingroups than outgroups. Moreover, this finding is in agreement with Wang and Li (2013), finding that anger mediated the link between social relations and rejection during ultimatum game. Thereby, strangers' violations cause third-party bystanders to feel more angry than classmates, resulting in harsher TPP.

In addition, perceived unfairness and anger can serve as a chain-mediating mechanism linking group membership and adolescents' TPP. The evaluation-emotional-behavioral model suggests that evaluating an event as unjust leads to anger toward the offender, which may then lead to punishment for violating social norms, even if punishment is costly (Seip et al., 2014). These results imply that ingroup violations induced stronger perceived unfairness than outgroup violations, resulting in a reduced level of anger and ultimately prompting individuals to exhibit a lower TPP. Our findings are consistent with previous research, finding that perceived unfairness and anger exerted a chain-mediating effect between fairness consideration and second-party punishment (Singer and Steinbeis, 2009; Mendoza et al., 2014). Consequently, identification with a classmate can influence an individual's perception and evaluation of an unfair event, subsequently impacting the level of anger experienced and ultimately altering the degree of TPP.

Finally, as previously reported among children (Wu and Gao, 2018), preliminary results indicated that boys perceive classmates' violation as less unfair, experience less anger, and impose softer punishments compared to strangers' violations, while girls only exhibit a small IGF on TPP. When gender was incorporated into the model, gender could negatively moderate the relationship between group membership and perceived unfairness. It supports the social role theory that boys have stronger IGF than girls (Eagly, 2009). This gender difference may be caused by different societal stereotypes and socialization processes for boys and girls (Rose and Rudolph, 2006). In adolescence, social norms expect boys' prescriptive roles to be agent, dominant, and assertive, while girls' prescriptive roles to be warm, communal and supportive (Koenig, 2018). Consequently, boys have stronger IGF than girls as a result of these experiences.

5 Implications of the study

To our knowledge, our research is the first to demonstrate IGF among adolescents' TPP. This finding has significant implications for the broader question of how morality is formed and developed. First, our results indicate that TPP is biased from childhood through adolescence and into adulthood, which completes the developmental trajectories associated with this bias. Second, an individual's perceived unfairness, anger and chain mediation between them may be a psychological mechanism contributing to this bias. Third, the indirect path of group membership and perceived unfairness is significant for boys, but not for girls, implying that gender modulates this indirect path. Using these results, we can better understand when and how group biases develop and who is more likely to exhibit them.

6 Limitations and future research

Like previous research, this study is subject to several limitations. Initially, the third-party punishment game used in our study involves costless self-reported punishment, which might be different from

incentivized punishment (Gummerum et al., 2016, 2020, 2022; Gonzalez-Gadea et al., 2022). Future studies should explore how TPP with real monetary incentives are affected by group membership. Furthermore, the identity of the classmate was not controlled. Different participants may have imagined different types of classmates and this could substantially increase the variance of classmate's IOS. This differentiation might substantially affect adolescents' interpersonal decision-making (Burnett Heyes et al., 2015), which needs to be strictly controlled in future studies. Finally, it is worth considering that various factors, such as compassion and social orientation value, could influence the association between group membership and TPP. Therefore, future research endeavors could benefit from the inclusion of additional variables in order to gain a more comprehensive understanding of this relationship.

7 Conclusion

Our results indicated that adolescents enacted more severe sanctions to stranger's violation than to classmate's violation during the third-party punishment task. Moreover, perceived unfairness and anger had a chain-mediating effect on the relationship between group membership and TPP. Additionally, the indirect path of group membership and perceived unfairness is significant for boys, but not for girls. These findings contribute to a deeper comprehension of the development mechanism of group bias in adolescents' TPP.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Faculty of Education at Henan Normal University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

CQ and ZZ designed the experiment. ML, QL, and CC collected and analyzed the data. ML and ZZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the National Natural Science Foundation of China [32000754], the Youth Foundation of the Ministry of Education of Humanities and Social Science Project of China [20YJC190030], the Philosophy and Social Science Foundation of Henan Province of China [2021CJY052], the Science and Technology Research Project of Henan Province [222102320386], and the Teacher Education Reform Project of Henan Province [2022-JSJYYB-011].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *J. Pers. Soc. Psychol.* 63, 596–612. doi: 10.1037/0022-3514.63.4.596

Balafoutas, L., and Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *Eur. Econ. Rev.* 56, 1773–1785. doi: 10.1016/j.euroecorev.2012.09.008

Balliet, D., Wu, J., and De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychol. Bull.* 140, 1556–1581. doi: 10.1037/a0037737

Bašić, Z., Falk, A., and Kosse, F. (2020). The development of egalitarian norm enforcement in childhood and adolescence. *J. Econ. Behav. Organ.* 179, 667–680. doi: 10.1016/j.jebo.2019.03.014

Baumgartner, T., Götte, L., Gügler, R., and Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Hum. Brain Mapp.* 33, 1452–1469. doi: 10.1002/hbm.21298

Bicskei, M., Lankau, M., and Bizer, K. (2016). Negative reciprocity and its relation to anger-like emotions in identity-homogeneous and-heterogeneous groups. *J. Econ. Psychol.* 54, 17–34. doi: 10.1016/j.jpubeco.2015.12.012

Blakemore, S. J. (2018). Avoiding social risk in adolescence. *Curr. Dir. Psychol. Sci.* 27, 116–122. doi: 10.1177/0963721417738144

Bonini, N., Hadjichristidis, C., Mazzocco, K., Demattè, M. L., Zampini, M., Sbarbati, A., et al. (2011). Pecunia olet: the role of incidental disgust in the ultimatum game. *Emotion* 11, 965–969. doi: 10.1037/a0022820

Buckholtz, J. W., and Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* 15, 655–661. doi: 10.1038/nn.3087

Burnett Heyes, S., Jih, Y. R., Block, P., Hiu, C. F., Holmes, E. A., and Lau, J. Y. (2015). Relationship reciprocation modulates resource allocation in adolescent social networks: developmental effects. *Child Dev.* 86, 1489–1506. doi: 10.1111/cdev.12396

Chawla, M., Earp, B. D., and Crockett, M. J. (2020). A neuroeconomic framework for investigating gender disparities in moralistic punishment. *Curr. Opin. Behav. Sci.* 34, 166–172. doi: 10.1016/j.cobeha.2020.03.011

Chen, S. P., and Bo, X. (2016). The influence of unfairness and punishment price to the demand of third-party punishment. *Stud. Psychol. Behav.* 14, 372–376. doi: 10.3969/j.issn.1672-0628.2016.03.013

Chierchia, G., Tufano, F., and Coricelli, G. (2020). The differential impact of friendship on cooperative and competitive coordination. *Theor. Decis.* 89, 423–452. doi: 10.1007/s11238-020-09763-3

Delton, A. W., and Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evol. Hum. Behav.* 38, 734–743. doi: 10.1016/j.evolhumbehav.2017.07.003

Du, K., Hunter, J. A., Scarf, D., and Ruffman, T. (2021). Chinese children's in-group favoritism is affected by age and gender. *J. Appl. Dev. Psychol.* 72, 101232. doi: 10.1016/j.appdev.2020.101232

Eagly, A. H. (2009). The his and hers of prosocial behavior: an examination of the social psychology of gender. *Am. Psychol.* 64, 644–658. doi: 10.1037/0003-066X.64.8.644

Fatfouta, R., Meshi, D., Merkl, A., and Heekeren, H. R. (2018). Accepting unfairness by a significant other is associated with reduced connectivity between medial prefrontal and dorsal anterior cingulate cortex. *Soc. Neurosci.* 13, 61–73. doi: 10.1080/17470919.2016.1252795

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Fehr, E., and Fischbacher, U. (2004). Third-party punishment and social norms. *Evol. Hum. Behav.* 25, 63–87. doi: 10.1016/S1090-5138(04)00005-4

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868. doi: 10.1162/003355399556151

Gonzalez-Gadea, M. L., Dominguez, A., and Petroni, A. (2022). Decisions and mechanisms of intergroup bias in children' third-party punishment. *Soc. Dev.* 31, 1194–1210. doi: 10.1111/sode.12608

Gummerum, M., and Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second-and third-party punishment behavior. *Cognition* 133, 97–103. doi: 10.1016/j.cognition.2014.06.001

Gummerum, M., López-Pérez, B., Van Dijk, E., and Van Dillen, L. F. (2020). When punishment is emotion-driven: Children's, adolescents', and adults' costly punishment of unfair allocations. *Soc. Dev.* 29, 126–142. doi: 10.1111/sode.12387

Gummerum, M., López-Pérez, B., Van Dijk, E., and Van Dillen, L. F. (2022). Ire and punishment: incidental anger and costly punishment in children, adolescents, and adults. *J. Exp. Child Psychol.* 218:105376. doi: 10.1016/j.jecp.2022.105376

Gummerum, M., Van Dillen, L. F., Van Dijk, E., and López-Pérez, B. (2016). Costly third-party interventions: the role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *J. Exp. Soc. Psychol.* 65, 94–104. doi: 10.1016/j.jesp.2016.04.004

Guo, R., Ding, J., and Wu, Z. (2020). How intergroup relation moderates group bias in third-party punishment. *Acta Psychol.* 205:103055. doi: 10.1016/j.actpsy.2020.103055

Guo, Z., Guo, R., Xu, C., and Wu, Z. (2022). Reflexive or reflective? Group bias in third-party punishment in Chinese and Western cultures. *J. Exp. Soc. Psychol.* 100:104284. doi: 10.1016/j.jesp.2022.104284

Guo, Q., Xu, P., Wu, R., and Hu, S. (2016). Effects of in-group favoritism and grade on the altruistic punishment behavior of primary school students. *Psychol. Dev. Educ.* 32, 402–408. doi: 10.16187/j.cnki.issn1001-4918.2016.04.03

Güth, W., and Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: motives, variations, and a survey of the recent literature. *J. Econ. Behav. Organ.* 108, 396–409. doi: 10.1016/j.jebo.2014.06.006

Hartsough, L. E., Ginther, M. R., and Marois, R. (2020). Distinct affective responses to second-and third-party norm violations. *Acta Psychol.* 205:103060. doi: 10.1016/j.actpsy.2020.103060

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science* 312, 1767–1770. doi: 10.1126/science.1127333

Henrich, J., and Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annu. Rev. Psychol.* 72, 207–240. doi: 10.1146/annurev-psych-081920-042106

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., et al. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proc. R. Soc. B: Biol. Sci.* 287:20192794. doi: 10.1098/rspb.2019.2794

Jin, Y., Gao, Q., Wang, Y., Xiao, L., Wu, M. S., and Zhou, Y. (2022). The perception-behavior dissociation in the ultimatum game in unmedicated patients with major depressive disorders. *J. Psychopathol. Clin. Sci.* 131, 253–264. doi: 10.1037/abn0000747

Jordan, J. J., McAuliffe, K., and Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proc. Natl. Acad. Sci. U. S. A.* 111, 12710–12715. doi: 10.1073/pnas.1402280111

Koenig, A. M. (2018). Comparing prescriptive and descriptive gender stereotypes about children, adults, and the elderly. *Front. Psychol.* 9:1086. doi: 10.3389/fpsyg.2018.01086

Kromer, E., and Bahçekapili, H. G. (2010). The influence of cooperative environment and gender on economic decisions in a third party punishment game. *Procedia. Soc. Behav. Sci.* 5, 250–254. doi: 10.1016/j.sbspro.2010.07.082

Krueger, F., and Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends Neurosci.* 39, 499–501. doi: 10.1016/j.tins.2016.06.004

Lazić, A., Purić, D., and Krstić, K. (2021). Does parochial cooperation exist in childhood and adolescence? A meta-analysis. *Int. J. Psychol.* 56, 917–933. doi: 10.1002/ijop.12791

Lee, Y. E., and Warneken, F. (2022). The influence of age and experience of (un) fairness on third-party punishment in children. *Soc. Dev.* 31, 1176–1193. doi: 10.1111/sode.12604

Lieberman, D., and Linke, L. (2007). The effect of social category on third party punishment. *Evol. Psychol.* 5, 147470490700500–147470490700305. doi: 10.1177/147470490700500203

Lu, T., and McKeown, S. (2018). The effects of empathy, perceived injustice and group identity on altruistic preferences: towards compensation or punishment. *J. Appl. Soc. Psychol.* 48, 683–691. doi: 10.1111/jasp.12558

Marshall, J., and McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nat. Rev. Psychol.* 1, 334–344. doi: 10.1038/s44159-022-00046-y

McAuliffe, K., Blake, P. R., Steinbeis, N., and Warneken, F. (2017). The developmental foundations of human fairness. *Nat. Hum. Behav.* 1:0042. doi: 10.1038/s41562-016-0042

McAuliffe, K., and Dunham, Y. (2016). Group bias in cooperative norm enforcement. *Philos. Trans. R. Soc. B* 371:20150073. doi: 10.1098/rstb.2015.0073

McAuliffe, K., Jordan, J. J., and Warneken, F. (2015). Costly third-party punishment in young children. *Cognition* 134, 1–10. doi: 10.1016/j.cognition.2014.08.013

McCall, C., Steinbeis, N., Ricard, M., and Singer, T. (2014). Compassion meditators show less anger, less punishment, and more compensation of victims in response to fairness violations. *Front. Behav. Neurosci.* 8:424. doi: 10.3389/fnbeh.2014.00424

Mendoza, S. A., Lane, S. P., and Amodio, D. M. (2014). For members only: ingroup punishment of fairness norm violations in the ultimatum game. *Soc. Psychol. Personal. Sci.* 5, 662–670. doi: 10.1177/1948550614527115

Pratto, F., Sidanius, J., Stallworth, L. M., and Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *J. Pers. Soc. Psychol.* 67, 741–763. doi: 10.1037/0022-3514.67.4.741

Riedl, K., Jensen, K., Call, J., and Tomasello, M. (2015). Restorative justice in children. *Curr. Biol.* 25, 1731–1735. doi: 10.1016/j.cub.2015.05.014

Rodriguez-Ruiz, C., Munoz-Reyes, J. A., Iglesias-Julios, M., Sanchez-Pages, S., and Turiegano, E. (2019). Sex affects the relationship between third party punishment and cooperation. *Sci. Rep.* 9:4288. doi: 10.1038/s41598-019-40909-8

Rose, A. J., and Rudolph, K. D. (2006). A review of sex differences in peer relationship processes: potential trade-offs for the emotional and behavioral development of girls and boys. *Psychol. Bull.* 132, 98–131. doi: 10.1037/0033-2909.132.1.98

Salali, G. D., Juda, M., and Henrich, J. (2015). Transmission and development of costly punishment in children. *Evol. Hum. Behav.* 36, 86–94. doi: 10.1016/j.evolhumbehav.2014.09.004

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., and Patton, G. C. (2018). The age of adolescence. *Lancet Child Adolesc.* 2, 223–228. doi: 10.1016/S2352-4642(18)30022-1

Seip, E. C., Van Dijk, W. W., and Rotteveel, M. (2014). Anger motivates costly punishment of unfair behavior. *Motiv. Emotion* 38, 578–588. doi: 10.1007/s11031-014-9395-4

Singer, T., and Steinbeis, N. (2009). Differential roles of fairness-and compassion-based motivations for cooperation, defection, and punishment. *Ann. N. Y. Acad. Sci.* 1167, 41–50. doi: 10.1111/j.1749-6632.2009.04733.x

Towner, E., Chierchia, G., and Blakemore, S. J. (2023). Sensitivity and specificity in affective and social learning in adolescence. *Trends Cogn. Sci.* 27, 642–655. doi: 10.1016/j.tics.2023.04.002

Wang, N., and Li, X. M. (2013). The mechanism and effects of "guanxi" on justice about unfair distribution: from the cold and hot perspectives. *Stud. Psychol. Behav.* 11, 239–244. doi: 10.3969/j.issn.1672-0628.2013.02.017

Wu, Z., and Gao, X. (2018). Preschoolers' group bias in punishing selfishness in the Ultimatum Game. *J. Exp. Child Psychol.* 166, 280–292. doi: 10.1016/j.jecp.2017.08.015

Yang, H., Zhang, Y., Lyu, Y., and Tang, C. (2023). Group bias under uncertain environment: A perspective of third-party punishment. *Acta. Psychol.* 237:103957. doi: 10.1016/j.actpsy.2023.103957

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., and Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *J. Exp. Psychol. Gen.* 145, 1448–1459. doi: 10.1037/xge0000190

Yudkin, D. A., Van Bavel, J. J., and Rhodes, M. (2019). Young children police group members at personal cost. *J. Exp. Psychol. Gen.* 149, 182–191. doi: 10.1037/xge0000613

Zhang, Z., Qi, C., Wang, Y., Zhao, H., Wang, X., and Gao, X. (2020). In-group favoritism or the black sheep effect? Group bias of fairness norm enforcement during economic games. *Adv. Psychol. Sci.* 28, 329–339. doi: 10.3724/SP.J.1042.2020.00329

Zhang, Z., Su, H., Li, M., Zhao, H., and Qi, C. (2022). Effects of ingroup identification on ingroup favouritism during fairness norm enforcement. *Behav. Sci.* 12:415. doi: 10.3390/bs12110415

Zhang, Z., Zhao, H., Liu, R., and Qi, C. (2021). Victim sensitivity and proposal size modulate the ingroup favoritism during fairness norm enforcement. *Front. Psychol.* 12:738447. doi: 10.3389/fpsyg.2021.738447

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 December 2023
DOI 10.3389/fpsyg.2023.1289209

OPEN ACCESS

EDITED BY
Carmelo Mario Vicario,
University of Messina, Italy

REVIEWED BY
Drew A. Curtis,
Angelo State University, United States
Christian L. Hart,
Texas Woman's University, United States

*CORRESPONDENCE
Enrique Armas-Vargas
✉ extearmasva@ull.edu.es

RECEIVED 05 September 2023
ACCEPTED 28 November 2023
PUBLISHED 20 December 2023

CITATION
Armas-Vargas E, Marrero RJ and Hernández-Cabrera JA (2023) Psychometric properties of the CEMA-A questionnaire: motives for lying.
Front. Psychol. 14:1289209.
doi: 10.3389/fpsyg.2023.1289209

Psychometric properties of the CEMA-A questionnaire: motives for lying

Enrique Armas-Vargas[1]*, Rosario J. Marrero[1,2] and Juan A. Hernández-Cabrera[1]

[1]Department of Clinical Psychology, Psychobiology and Methodology, Universidad de La Laguna, Tenerife, Spain, [2]Instituto Universitario de Neurociencia (IUNE), Universidad de La Laguna, Tenerife, Spain

Previous research on the motives for lying lacks factorial models that allow grouping of motives into specific categories. The objective of this study is to confirm the factorial structure of the questionnaire of motives for lying (CEMA-A). Participants were 1,722 adults residing in the Canary Islands (Spain) who completed the CEMA-A and the Eysenck Personality Questionnaire (EPQ-R). The four-dimensional structure of the questionnaire was confirmed ($\chi^2 = 1460.97$, df = 325, $p = 0.001$; CFI = 0.94; TLI = 0.93; NFI = 0.93; NNFI = 0.93; RMSEA = 0.05, CI = 0.051–0.057; SRMR = 0.04). The four factors of the CEMA-A were Intrapersonal Motivation–Emotionality, Interpersonal Motivation–Sociability, Egoism/Hardness Motivation, and Malicious Motivation, with an internal consistency between 0.79 and 0.91. Invariance analyses confirmed the equivalence of the instrument for men and women. The CEMA-A factors positively correlated with Neuroticism and Psychoticism, and negatively with Dissimulation. Extraversion was not related to any of the factors, and only displayed a low negative correlation with Intrapersonal Motivation–Emotionality. Analysis of variance showed that men scored higher in Egoism/Hardness and Malicious Motivation. The CEMA-A has proven capable of apprehending the motives for lying and has adequate psychometric criteria for use in various populations.

KEYWORDS

motives for lying, intrapersonal motivation-emotionality, interpersonal motivation-sociability, egoism/hardness motivation, malicious motivation

1 Introduction

A lie is a multidimensional construct (DePaulo et al., 1996; Phillips et al., 2011; Muzinic et al., 2016), defined as a form of verbal deception, where there is a deliberate attempt to hide, falsify, generate and/or manipulate, in some way, factual, and/or emotional information, to encourage in the other a belief that the communicator themself considers false (Knapp and Comadena, 1979; Ekman, 1985/2001; Miller and Stiff, 1993; Buller et al., 1994; Masip et al., 2004; Vrij, 2008). People evaluate lying from two positions, by assigning a negative image to those who lie and by rationalizing or justifying the lie when it is used by the individual themself (Nyberg, 1993; Kashy and DePaulo, 1996; Bond and DePaulo, 2006). Thus, more intentionality is attributed, and the label of liar is assigned more to others than to oneself when lying (Curtis, 2021). Research suggests that people view their everyday lives as small, and unimportant, rarely plan them, and unconcerned about being discovered (DePaulo et al., 2004; Bond and DePaulo, 2006). Most lies that are considered serious are motivated by the desire to cover up a personal

fault, a discredited fact or to hide transgressions that, if discovered, could have serious consequences for the identity and reputation of the liar (McCornack and Levine, 1990; Metts, 1994; DePaulo et al., 2003a, 2004). These types of lies are more carefully planned, and are often unjustifiable, immoral, or illegal (DePaulo et al., 2004). Therefore, unless there is a psychopathological problem (Curtis and Hart, 2022), people often use deception, when telling the truth is a problem (McCornack et al., 2014; Levine et al., 2016; Moshagen et al., 2020). Since lying is intentional, that people lie for a reason or motive is implicit (Bond and DePaulo, 2006; Levine et al., 2010), lying, in itself, is not a goal, but a means to achieve another (Levine et al., 2010). For example, someone tells their partner that they are at home (a lie) when they are in fact with a lover. This lie does not seek to convince the partner of their where about, since they could have excused themselves in another way, to convince them of their fidelity (the goal).

In general, people may tell a lie at some point, despite it being considered a reprehensible act with harmful consequences (Bok, 1978; Teasdale and Kent, 1995; Solomon, 2009; Curtis and Hart, 2015). However, lying every day is not common behavior for most people (Serota et al., 2010, 2022; Serota and Levine, 2015). Studies in the field of everyday lies find that people report an average of one to two lies a day (DePaulo and Kashy, 1998; Serota et al., 2010; Serota and Levine, 2015). However, the average may be distorted by extreme scores from people who often lie. These differences in the frequency of lying may also be related to sociodemographic variables. Some studies suggest that young people and men admit to lying more often (DePaulo et al., 1996; Serota et al., 2010; Armas-Vargas, 2017a), although the difference in frequency of lying between men and women is very small (Gerlach et al., 2019). Other research has found that gender differences vary depending on the subject matter of the lie (DePaulo et al., 1996; Feldman et al., 2002; Haselton et al., 2005; Erat and Gneezy, 2012). Various studies suggest that lying decreases with age (Jensen et al., 2004; Serota et al., 2010). Adolescents tend to lie more than university students, who do so less than the general adult population (DePaulo et al., 1996; Serota et al., 2010; Levine et al., 2013; Armas-Vargas, 2020, 2021b).

Most research has been carried out in the area of lie detection (Vrij and Ganis, 2014). The truth-default theory (Levine, 2014; Levine et al., 2022) is one of the most widely accepted theories about human deception detection. This theory proposes that people tell the truth by default, that is, they are honest most of the time, and are more likely to believe that others tell them the truth rather than lies. Thus, people do not usually lie except when the truth is an obstacle to goal attainment (Levine et al., 2010). However, if a situation becomes problematic, people can then lie. On the other hand, the self-concept maintenance theory (Mazar et al., 2008; Ariely, 2012) proposes that people are more likely to lie when the ego is depleted (Mead et al., 2009). Therefore, the aim of a person who deceives is to satisfy complex intrinsic motivations, such as maintaining a favorable self-concept (Mazar et al., 2008). Similarly, from a self-presentational perspective, DePaulo et al. (2003b) propose that people mainly lie for psychological reasons to protect or to give a better image of themselves, that is, to deliberately try to manage others' impressions of them. Furthermore, DePaulo et al. (2003b) suggest that deception and truth can be distributed along a continuum rather than considered different dimensions. The reasons for lying or telling the truth are the same: people are interested in giving a good image or describing important aspects of themselves. However, self-presentation is not the only reason why one can lie. In a transcultural study, Levine et al. (2016) found that there were different types of deception motives such as maintaining a positive self-image, protecting others, avoiding others, seeking an advantage, social politeness, hiding a transgression, being malicious, and joking.

Furthermore, some people who lie give socially desirable responses and misrepresent their motivations for lying (DePaulo et al., 2003b). Much research indicates that people offer a positive view of themselves, highlighting positive features such as that they are better, more honest, and more moral than others (Alicke et al., 1995). These beliefs are identified with self-deception. For self-deception to produce positive effects on the person, individuals must, by definition, be unaware of its illusory basis (Baumeister, 1993). According to Trivers (2002), "the hallmark of self-deception in the service of deceit is the denial of deception, the unconscious running of selfish and deceitful ploys, the creation of a public person as an altruist and a person beneffective in the lives of others" (p. 276). Therefore, as they are not fully aware of their motivations for lying, these people can confidently and "honestly" claim that their lies were altruistically motivated. However, according to cognitive dissonance theory (Festinger, 1957), altruistic interpretations of deception may not completely dispel the dissonance of the person lying. In these cases, the person may wield feeling guilty about the lie as a way to reduce cognitive dissonance. When they express guilt for lying to others, they are reinforcing their positive view of themselves. Individuals who feel and express guilt for their misdeeds are often considered better people than those who show no remorse (Baumeister, 1997).

Other authors have attempted to capture and classify the motives for lying. Turner et al. (1975) list five motivations for lying: (a) to save face (to protect identity, self-esteem), (b) to manage or handle relationships (to end a relationship), (c) to exploit others (by manipulating, having control, power, and influence over the other), (d) to avoid tensions or conflicts (controlling a conversation to avoid it being uncomfortable or triggering an argument) and (e) to control situations (to maintain, redirect or end interaction with the other). Buller and Burgoon (1996) point out that lying is employed for three main reasons/motives: (a) "instrumental" (to gain power, influence others, avoid disapproval, or do harm), (b) "identity" (to improve the image we present to others, avoid shame, improve or protect self-esteem, and increase social desirability), and (c) "relational" (to influence our relationships with others).

Another proposed categorization of the motives for lying is based on (a) whether the liar is "centered on themselves" (egotistical, to protect themselves) or on the other person (to protect others), and (b) whether the liar is "altruistically" or "maliciously" motivated (DePaulo et al., 1996, 2003a; DePaulo and Kashy, 1998; Vrij, 2000). Altruistic lies also allow one to protect one's well-being (Ennis et al., 2008) and have been classified as considerably more acceptable than egotistical lies (for one's own benefit or for malicious purposes) (Lindskold and Walters, 1983; Seiter et al., 2002). DePaulo et al. (1996) found that people lie far more about themselves than they do about others. The motives behind the lies were mostly selfish, and many more lies were told for emotional reasons (to protect themselves from shame, or their own feelings) than for personal advantage (to obtain benefits or material gain).

From a qualitative perspective, the motivations for lying have been classified from two dimensions: protective versus beneficial lies and self-oriented versus other-oriented lies (Arcimowicz et al.,

2015). The combination of both dimensions facilitates the identification of four types of lies: egoistic (self-oriented/beneficial), self-defensive (self-oriented/protective), pleasing (other-oriented/beneficial), and sheltering (other-oriented/protective). Some egoistic lies cited in the interviews were for material gains or admiration from others. Self-defensive lies included avoiding responsibility, discussions, or negative consequences. Pleasing lies were related to making someone happy, and sheltering lies were associated with protecting someone from distress or avoiding hurting someone else. These last two categories were more difficult to distinguish. In general, people lie primarily for protective motivations that allow them to avoid punishment rather than for personal benefits.

The role played by inter-individual differences may affect the probability of lying (McLeod and Genereux, 2008), as well as the different motives for lying and achieving certain goals or desires (Buller and Burgoon, 1996; Olson and Weber, 2004). Some studies point out the importance of personality traits in the probability of and motives for lying (McArthur et al., 2022). Machiavellianism or extraversion are associated with frequency and different types of lying (Kashy and DePaulo, 1996; McLeod and Genereux, 2008; Hart et al., 2019). In the prison population, lying has been found to be associated with both neuroticism and psychoticism (Gudjonsson and Sigurdsson, 2004). Fullam et al. (2009) found that people with a high level of psychoticism showed a low level of conditioning to social norms, a low level of fear and avoidance of harm, and were more likely to lie. The results of the study revealed the importance of analyzing the role of the traits of insensitivity and emotional deficit (typical of psychoticism and neuroticism) in the tasks that evaluate the cognitive elements that may be involved in deceiving and manipulating others. Giammarco et al. (2013) also found an association between greater ability to deceive and the Dark Triad of Personality (Machiavellianism, psychopathy, and narcissism). Furthermore, an important motivator in lying is emotions (Ekman, 1985/2001). Lying is mainly motivated by negative emotions, such as anxiety, fear (Ekman, 1985/2001; Tangney et al., 1996), or guilt, which arises when there is a discrepancy between internalized values and actual behavior (Mosher, 1968; Ekman, 1985/2001; Millar and Tesser, 1988); shame, when a person does not meet their own personal moral standards (Keltner and Buswell, 1996; Tangney et al., 1996; DePaulo et al., 2003a); and insecurity, fear of rejection and criticism (Armas-Vargas, 2021a,b). People are motivated to lie mainly through certain emotional needs, which are satisfied through social interaction, the instrumentalization of relationships, or harming others (Armas-Vargas, 2021a). That is, personal/emotional motives may be based on other more social, instrumental/selfish, or malicious motives (Armas-Vargas, 2021a). Many of these emotional motives may be implicit or escape awareness (McClelland et al., 1989; Bargh, 1990; Bargh and Chartrand, 1999; Bargh et al., 2001; Custers and Aarts, 2005), while interpersonal, instrumental, and malicious motives imply heightened awareness (Schooler and Schreiber, 2005; Touré-Tillery and Fishbach, 2014).

Several studies have tried to classify the motives for deception using different methods, such as researchers' expert judgment, literature reviews, and analysis of diary records and, interviews, or surveys (Turner et al., 1975; Ekman et al., 1989; DePaulo et al., 1996; Kashy and DePaulo, 1996; McLeod and Genereux, 2008; Phillips et al., 2011; Arcimowicz et al., 2015; Levine et al., 2016). However, few studies have designed self-report instruments to identify and categorize motives using factor analysis. One of the self-report instruments proposed, designed by Hart et al. (2019), identifies two categories that evaluate relational and antisocial motives. In a later study, Hart et al. (2020) found three categories of motives for lying: self-serving lies (such as avoiding the consequences of bad behavior and self-promotion), altruistic or benevolent lies (to benefit another), and vindictive lies to harm another person.

The aim of this study is to analyze the psychometric properties of an instrument that assesses people's main motives for lying in their daily lives. The instrument was constructed to combine the different theoretical models described, as well as other typologies proposed by various authors on the motives for lying. The CEMA-A questionnaire was based on a review of the literature, to integrate the various motives behind every day lies. The instrument design mainly took into account the role of emotions in lying (Ekman, 1985/2001; Tangney et al., 1996); the five motivations for lying proposed by Turner et al. (1975); the three main reasons/motives of Buller and Burgoon (1996); the 10 pancultural deception motives of Levine et al. (2016); the research on self-presentational motives for lying in everyday life (DePaulo et al., 1996, 2003a; Kashy and DePaulo, 1996; DePaulo and Kashy, 1998), and personality variables related to lying (Olson and Weber, 2004; McLeod and Genereux, 2008; Armas-Vargas, 2017a; Armas-Vargas, 2020; Armas-Vargas, 2021b). In a pilot study (Armas-Vargas, 2021a), a four-factor structure was obtained, after exploratory factor analysis (EFA). The "Intrapersonal Motivation–Emotionality" category evaluates motives related to self-deception and negative emotions (shame, insecurity, fear of rejection and criticism). "Interpersonal Motivation–Sociability" evaluates reasons for the benefit of social relationships (to excuse or justify oneself, avoid conflicts with others, and for reasons of a prosocial nature). "Egoism/Hardness Motivation" measures motives related to using relationships for one's own benefit (to obtain advantage, manipulate others, present a good image and impress others). And finally, the "Malicious Motivation" category evaluates motives related to covert or direct harm, or false accusations that cause harm (Armas-Vargas, 2021a). Unlike the test proposed by Hart et al. (2019), the CEMA-A posits two new categories: Intrapersonal Motivation and Egoism/Hardness Motivation. The other two factors of relational and antisocial motives proposed by Hart et al. (2019) correspond, to a certain extent, with Interpersonal Motivation–Sociability and Malicious Motivation, respectively.

The objective of this work is to study the psychometric properties of the CEMA-A instrument. Specifically, it will analyze whether the factorial structure found in the previous exploratory analyses (Armas-Vargas, 2021a), remains stable. Next, confirmatory factor analysis (CFA) will be used to check construct validity whether the data conform to the proposed four-factor structure. The internal consistency of the four scales and the total test will be studied, along with the temporal stability provided by the test–retest correlations of the factors. Likewise, factorial invariance will be examined to verify whether the structure is similar between men and women. Convergent and discriminant validity will be checked by analyzing the relationship with other personality variables. Finally, the mean differences of the various factors of the CEMA-A will be analyzed according to gender and level of education.

2 Materials and methods

2.1 Participants

The total sample was 1,722 adults (Sample 3) from the general population of the Canary Islands (Spain), aged 18 to 77 years (Mage = 35.13, SD = 13.74): 55.89% women (N = 962) and 44.11% men (N = 760). The total sample was divided into subsamples for the different phases of the study. Sample 1 consisted of 520 participants aged 18 to 76 years (Mage = 36.80, SD = 14.44) and was used to perform the EFA. Sample 2 consisted of 1,202 participants aged 18 to 77 years (Mage = 34.41, SD = 13.37), and was used for CFA and analysis of invariance, based on gender. Sample 3 was used to perform analysis of variance (MANOVA) based on gender and level of education. Sample 4 consisted of 529 participants from the total sample, aged 18 to 71 years (Mage = 34.90, SD = 13.25), selected to analyze the temporal stability of the factors. Table 1 presents the characteristics of the total sample and the different subsamples.

2.2 Instruments

Questionnaire for the Evaluation of Deceit, Lies and Self-deception (CEMA) (Armas-Vargas, 2021a). The instrument was developed based on Muñiz and Fonseca-Pedrero (2019) recommendations for test construction. This self-report instrument designed to assess variables associated with "deceit, lying, concealment, and self-deception" consists of four sub-questionnaires: The Motives for Lying (CEMA-A); Opinions about Self-Deception Lying (CEMA-B); Content of Lies (CEMA-C); and Receivers of the Lies (CEMA-D). In this study, we validate the CEMA-A subquestionnaire that assesses people's motives for lying in their daily lives. Questionnaire development drew from a pool of 80 items related to personal–emotional variables (associated with protection of the self, such as fear of rejection, fear of what others will say, insecurity, self-esteem problems, self-deception); items related to instrumental content, manipulation of others, pro-image, and self-presentation (more selfish, intention to benefit oneself); other items concerning lies in social interactions (lies that are altruistic, prosocial, or beneficial to others); and finally, items related to malice or harming others. Two independent experts checked the wording and clarity of the items; when they disagreed, a third expert was consulted. Participants were informed that the aim of the study was to investigate the motives people may have for lying. Specifically, participants received the information that "lying includes both deliberately omitting relevant information and telling someone something that is not true." Then, to minimize problems of social desirability, participants were also told that it is normal to lie from time to time and the fact of being able to lie is not censored, but research is interested in studying the reasons why one might lie at some point. Finally, participants were asked to indicate the reasons or motives for which they usually deceive, lie, or withhold information from others and to indicate on a Likert-type scale of seven alternatives (1 = rarely, 2 = from time to time, 3 = sometimes, 4 = usually, 5 = very often, 6 = many times, and 7 = always), which of the listed motives they generally use to a greater or lesser extent. They were thanked for their participation and asked to be honest in their answers.

In the previous pilot study (Armas-Vargas, 2021a), an exploratory factor analysis (oblimin rotation) was applied. Items that saturated on two factors and items with factor loadings below 0.40 were eliminated from the factor analysis, reducing the number of items from 80 to 45. The CEMA-A questionnaire was finally composed of 45 items, and a factorial structure of four factors or general categories was obtained: Intrapersonal Motivation–Emotionality, Interpersonal Motivation–Sociability, Egoism/Hardness Motivation, and Malicious Motivation. The Intrapersonal Motivation–Emotionality category evaluates motives related to self-deception and negative emotions; Interpersonal Motivation–Sociability collects motives related to maintaining positive social relationships; Egoism/Hardness Motivation measures motives related to using relationships for one's own benefit; and the Malicious Motivation category evaluates motives related to covert or direct harm, or false accusations that cause harm (Armas-Vargas, 2021a). Interpersonal Motivation–Sociability and Egoism/Hardness Motivation both refer to the domain of interpersonal relationships. However, in the Egoism/Hardness motives, the intention of the individual who lies is to benefit him/

TABLE 1 Sociodemographic characteristics of the participants.

	Sample 1 (n = 520)	Sample 2 (n = 1,202)	Sample 3 (n = 1722)	Sample 4 (n = 529)
Sex (Women, Men) (%)	(55.77/44.23)	(55.95/44.05)	(55.89/44.11)	(50.28/49.72)
Age (M, SD)	36.80 (14.44)	34.41 (13.37)	35.13 (13.74)	34.90 (13.25)
Civil Status (%)				
Single	65.31	68.55	67.57	63.33
Married	25.97	22.71	23.70	25.74
Separated	3.10	2.10	2.40	8.35
Divorced	5.62	6.64	6.33	2.58
Level of education (%)				
Primary	3.08	4.49	4.06	5.29
Secondary	14.23	13.81	13.94	16.24
Baccalaureate/Technical studies	38.65	44.43	42.69	55.78
University	44.04	37.27	39.31	22.69

herself with the act of lying, whereas in the Interpersonal Motivation–Sociability, the intention of the individual is more prosocial: he/she intends to benefit others with the act of lying. On the other hand, the Intrapersonal Motivation–Emotionality factor is related to more personal motivations, where the person "avoids or does not want to face the truth and reality," indirectly obtaining a "self-benefit, without instrumentalizing anyone" by avoiding facing reality. In the Egoism/Hardness Motivation factor, the person intends to gain self-benefit by "manipulating and instrumentalizing others." In this second case, the person acts and confronts reality in order to achieve a certain goal. The total reliability of Cronbach's alpha was 0.97 and the omega coefficient ωj = 0.79.

Eysenck Personality Questionnaire – Revised (EPQ-R; Eysenck and Eysenck, 1997). It explores three personality traits: (1) Extraversion (sociable, active, assertive, sensation-seeking); (2) Neuroticism (anxious, depressed, guilt); and (3) Psychoticism (aggressive, cold, egocentric, impulsive, antisocial). It also includes the Lie scale, intended to measure the tendencies of examinees to "fake good" when they complete the questionnaire. It is made up of 83 items with two response alternatives (true or false), referring to the person's way of acting, feeling and thinking. Because it is a shorter tool, the EPQ-R was used in this study to assess the personality characteristics that have been linked to lying, such as psychoticism, neuroticism, and extraversion. Since no other tests of motives for lying have been validated in Spanish, the EPQ-R was used to assess convergent validity through the lie scale, along with discriminant validity, to distinguish between motives for lying and personality traits that have previously been weakly correlated (Gudjonsson and Sigurdsson, 2004; McLeod and Genereux, 2008; Hart et al., 2019). Internal consistency oscillates between 0.71 and 0.86.

2.3 Procedure

Data collection was done by fourth-year psychology undergraduates and master's students of general health psychology at the University of La Laguna for three academic years 2020–2023. This study was not preregistered. Samples 1 ($N = 520$) and 2 ($N = 1{,}202$) were obtained in 2020, and 2021 and 2023, respectively. Sample 3 ($N = 1722$) is the sum of both samples, and Sample 4 ($N = 529$) was randomly drawn from the whole sample. The students were trained to administer the aforementioned tests, order to play the role of evaluators. Sampling was incidental for convenience (Gil-Escudero and Martínez-Arias, 2001). The students had to select 15 to 20 people from their close environment, homogenized by gender, to whom they would apply the instrument. They were informed about the objective of the study, voluntarily accepted to collaborate, and gave their written informed consent. Participants received an envelope containing an identification code and tests. One week later, the sealed envelope was collected, to guarantee anonymity. Participants were instructed to write a contact telephone number on the envelope, so that they could be contacted for a second retest. After four weeks, half of the sample of 1,200 was randomly selected and, of the 600 participants selected, 529 had returned the envelope with the retest completed. The participants completed the questionnaires independently, at home and on paper in approximately 30 min. No reward was offered for participation. The study was carried out in accordance with the Declaration of Helsinki and was approved by the Research Ethics and Animal Welfare Committee of the University of La Laguna (Registration Number: CEIBA2023-3299).

2.4 Data analysis

The data were analyzed using R version 4.0.5 (R Core Team, 2017), the Lavaan package (Rosseel, 2012), and the syntax described by ULLRToolbox (Hernández and Betancort, 2018). Initially, an EFA was performed with Sample 1 ($N = 520$). This sample was used to verify whether the same four-factor structure remained stable with 45 items proposed by Armas-Vargas (2021a). The procedure used to determine the number of factors was the optimal application of Horn's parallel analysis (Timmerman and Lorenzo-Seva, 2011). An EFA was performed on principal axes and oblique rotation (oblimin) since a correlation between the factors was expected.

Secondly, CFA was performed with 1,202 participants (Sample 2). The objective was to check the factorial structure of the questionnaire using the four-factor model obtained previously. The model fit was estimated using the maximum likelihood estimation method (Brown, 2006) was verified using a comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR), reported in the bibliography as adequate for ordinal data (Abad et al., 2011; Byrne, 2012). The expected values for an acceptable fit were around 0.90 for the CFI, TLI, normed fit (NFI) and non-normed fit (NNFI) indices (Kline, 2011). Values under 0.05 for SRMR and under 0.10 for RMSEA, with a 90% confidence interval, indicate reasonable model fit (Browne and Cudeck, 1989; MacCallum et al., 1996). To statistically compare the four-dimensional model, we used the χ^2 difference test. The reliability of the CEMA-A was evaluated using omega coefficient (McDonald, 1999). The omega coefficient (ω) is more precise than Cronbach's alpha (α) because reliability can be directly calculated using the estimates of the CFA parameters, resulting in much greater stability when dealing with non-continuous data (Gadermann et al., 2012; Dunn et al., 2014).

Thirdly, the factorial invariance of the CEMA-A based on gender was analyzed with the 1,202 participants (Sample 2), using the multigroup CFA. The configural invariance test shows whether the same items are associated with the same construct. After checking the configural invariance we tested the metric invariance by restricting the factorial loadings of similar items, so that they were the same in the different groups. To determine the metric invariance of the groups, we performed a $\Delta\chi^2$ test (Sass, 2011). If the metric model does not differ from the configural model, the metric invariance is inferred.

Fourthly, to analyze convergent and discriminant validity, Sample 1 participants completed the EPQ-R questionnaire. The association between the CEMA-A and the EPQ-R scales was analyzed using Pearson correlation.

Fifthly, with 1,722 participants (Sample 3), we analyzed the mean differences of the different factors of the CEMA-A by MANOVA, according to gender and level of education. The MANOVA effect size was estimated using partial η2, considering 0.01 as small, 0.06 as medium and 0.14 as large.

Finally, we used the test–retest method (Aldridge et al., 2017) to analyze the stability of the CEMA-A (Sample 4), after four weeks. Vuong's (1989) test was used to compare the predicted probabilities of non-nested models. First, it allows us to check whether two models

are distinguishable, and then, to determines whether the second model shows a better fit than the first. Under the premise of the null hypothesis, it is proposed that the two non-nested models fit equally well, that is, the expected value of their log-likelihood coefficient is equal to zero.

3 Results

3.1 Exploratory factor analysis

Sample 1 ($N = 529$) was used to verify that the properties of the data were adequate to perform EFA. The Kaiser–Meyer–Olkin index (KMO = 0.96) and Bartlett's test of sphericity were significant ($\chi^2(990) = 16{,}501$; $p < 0.001$), indicating that the analysis was feasible. The parallel analysis method (Horn, 1965) was used to decide the number of factors to extract. The scree test is a graphical representation of the magnitude of the eigenvalues and helps to identify the optimal number of factors that should be extracted. The scree test yielded only four factors that were included in the final scale (Figure 1). EFA (Sample 1) showed a four-factor structure with 43 items that explained 54.35% of the total variance (Intrapersonal Motivation, 18.37%; Egoism/Hardness Motivation, 15.45%; Interpersonal Motivation, 14.77%; and Malicious Motivation, 5.75%). Table 2 shows the standard deviation, skewness, kurtosis, and factor loading of each item. Of the 45 original CEMA-A items, two were deleted (item 23: "To feign a life I do not have"; item 32: "Out of jealousy") because their means were too low and produced a floor effect. Table 3 shows the eigenvalue, explained and cumulative variance, as well as the Cronbach's alpha and Omega hierarchical reliability of the four factors of the CEMA-A with the 43 items.

Next, a first-order EFA was applied again. Items whose factor loading was <0.40 and those that saturated in two factors were eliminated (≥ 0.30). Based on these criteria, the following items were eliminated: 7, 9, 13 and 39 of the Intrapersonal Motivation factor; items 23, 28, 32, 37, and 45 of the Egoism/Hardness Motivation factor; items 3, 11, 15, 21, 26, and 41 of the Interpersonal Motivation factor; and items 4 and 25 of the Malicious Motivation factor. With the 28 items, the KMO index was 0.95 and Bartlett's sphericity test was again significant ($\chi^2(378) = 10{,}026$; $p < 0.001$). The four-factor structure was maintained with the 28 items. Internal consistency was calculated using Cronbach's alpha and Hierarchical Omega, which were 0.95 and 0.77, respectively, for the total scale. The Intrapersonal Motivation factor showed $\alpha = 0.92$ and $\omega j = 0.72$; the Egoism/Hardness Motivation factor, $\alpha = 0.93$ and $\omega j = 0.83$; the Interpersonal Motivation factor, $\alpha = 0.89$ and $\omega j = 0.77$; and for the Malicious Motivation factor it was $\alpha = 0.77$ and $\omega j = 0.72$. Of the final structure of 28 items, the factors for Intrapersonal Motivation, Egoism/Hardness, Interpersonal Motivation, and Malicious Motivation explained 18.74, 16.93, 15.84, and 6.59% of the total variance, respectively. As can be seen, the correlation between the different factors was high, mainly between Intrapersonal Motivation and Egoism/Hardness Motivation ($r = 0.70$) (Table 4).

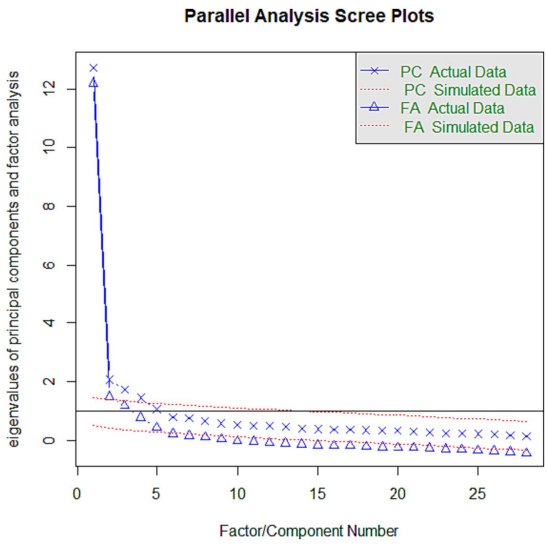


FIGURE 1
Parallel analysis scree plots.

3.2 Confirmatory factor analysis

To study the dimensional structure of the scale, we performed CFA with Sample 2, based on the model obtained with Sample 1. To analyze construct validity, we used a four-factor model with the 28 items, using the maximum likelihood estimation method. Figure 1 displays the results of the CFA of the four-factor model. To better evaluate the model parameters, taking into account the recommendations of other authors (Brown, 2015), we considered several indices simultaneously. Figure 1 shows the best fit model and normalized path coefficients for each variable observed. All item loadings were found to be at an acceptable level (≥ 0.47), and all parameter estimates were significantly different from 0. Latent correlation indices between model factors were high, for example, the latent correlation between the Egoism/Hardness and Malicious Motivation factors was r = 0.81.

When the proposed theoretical model was tested (Figure 2), an adequate fit to the data was obtained (Table 5). Applying the good fit statistics in this model resulted in the following: ($\chi^2 = 1{,}460.97$, df = 325, $p < 0.001$; CFI = 0.94; TLI = 0.93; NFI = 0.93; NNFI = 0.93; RMSEA = 0.05, CI = 0.051–0.057; SRMR = 0.04). It should be noted that all the parameters indicated in Figure 2 (factorial loadings, correlation between factors and measurement errors of the items) were significant for $p < 0.001$. Internal consistency was calculated using the McDonald omega coefficient for four factors. The Intrapersonal motivation factor presented $\omega = 0.91$, the Egoism/Hardness motivation factor, $\omega = 0.88$, the Interpersonal motivation factor, $\omega = 0.84$, and the Malicious motivation factor, $\omega = 0.79$.

3.3 Invariance of the CEMA-A factorial structure

Multigroup Confirmatory Factor Analysis. To check whether the factorial structure was similar according to gender (configural invariance), the parameters were estimated simultaneously for each gender level. The multigroup CFA fit indices were ($\chi^2 = 2082.76$, df = 650, $p < 0.001$; CFI = 0.93; TLI = 0.92; NFI = 0.90; NNFI = 0.92;

10.3389/fpsyg.2023.1289209

TABLE 2 Mean (*M*), standard deviation (*SD*), skewness, kurtosis and factor loading for CEMA-A (43 items).

Factor loading CEMA-A								
Reagents	*M*	*SD*	Skewness	Kurtosis	F1	F2	F3	F4
19. For fear of facing reality.	1.88	1.21	1.71	3.01	0.85			
17. Not to face the truth.	1.85	1.25	1.88	3.68	0.85			
10. Because I do not accept myself as I am.	1.65	1.14	2.11	4.60	0.78			
22. Because I feel insecure.	1.99	1.31	1.58	2.29	0.72			
16. Not to reveal my own meanness.	1.85	1.21	1.89	4.11	0.71			
30. Because it's hard for me to accept things as they are.	1.71	1.10	1.85	3.51	0.69			
24. For fear of what they will say.	2.22	1.39	1.08	0.61	0.60			
38. Out of shame to admit the truth.	1.97	1.28	1.57	2.25	0.59			
13. To be accepted by others.	1.90	1.31	1.60	2.06	0.39			
39. Because "telling the truth" hurts more.	2.12	1.41	1.50	1.83	0.33			
7. Because it is easier for me to lie than tell the truth.	1.88	1.27	1.63	2.43	0.28			
9. Due to mistrust.	2.25	1.38	1.37	1.68	0.27			
29.To get an advantage over others.	1.59	1.03	2.28	6.08		0.91		
6. To try to win an argument with someone.	1.73	1.17	1.91	3.75		0.83		
44. Because it is easier to manipulate others.	1.51	1.03	2.70	8.28		0.79		
36. To benefit from something.	1.98	1.35	1.56	2.12	−0.26	0.76		
12. To get what I want.	2.04	1.35	1.55	2.04		0.68		
5. To impress others.	1.88	1.32	1.89	3.50		0.63		
42.To earn the respect and admiration of others.	1.63	1.19	2.34	5.68		0.62		
27. To give a better image of myself.	1.99	1.30	1.45	1.69		0.43		
28. To seek the approval of others.	1.58	1.02	2.13	4.89		0.38		
45. Because it helps me to relate.	1.66	1.12	2.07	4.41	0.27	0.38		
37. Because I cannot help it.	1.47	0.97	2.67	8.53		0.34		
18. To avoid problems with others.	2.78	1.42	0.91	0.55			0.77	
33. To avoid having to explain.	2.72	1.47	1.07	0.83			0.74	
35. To make others feel good.	2.76	1.55	0.91	0.22			0.71	
20.To hide certain information.	2.66	1.43	1.15	1.12			0.69	
2. So as not to offend others.	3.35	1.52	0.50	−0.44			0.62	
34. To hide something I know is wrong.	2.44	1.36	1.10	1.00			0.57	
43.To be kind and cordial to others.	2.51	1.40	0.96	0.55			0.54	
14. For fear of punishment.	2.27	1.33	1.26	1.68			0.47	
11. To hide certain problems or difficulties.	2.40	1.38	1.15	1.06	0.35		0.44	
15. To protect myself.	2.38	1.43	1.11	0.68	0.27		0.36	
21. To defend myself against the attacks of others.	2.10	1.34	1.30	1.18		0.29	0.30	
3. To save face.	2.56	1.49	0.87	0.21			0.29	
26.To avoid telling or acknowledging the truth.	1.94	1.20	1.77	3.77			0.28	
41.To avoid taking responsibility for something.	2.09	1.25	1.25	1.48			0.27	
25. To give a bad image of another person.	2.02	1.39	1.47	1.73			0.34	0.69
8. To give false information about another person.	1.69	1.22	2.12	4.43				0.59
4. Not to make others feel bad.	2.59	1.58	0.91	0.03			0.39	−0.58
40. To falsely accuse another person and cause them harm.	1.41	0.89	2.53	6.68	0.28			0.49
1. To raise doubts about another person.	1.88	1.37	1.65	2.03				0.49
31.To make the other feel guilty.	1.57	1.11	2.38	6.08				0.48

F1, Intrapersonal motivation; F2, Egoism/Hardness motivation; F3, Interpersonal motivation; F4, Malicious motivation.

TABLE 3 Factor analysis of the CEMA-A questionnaire ($N = 520$).

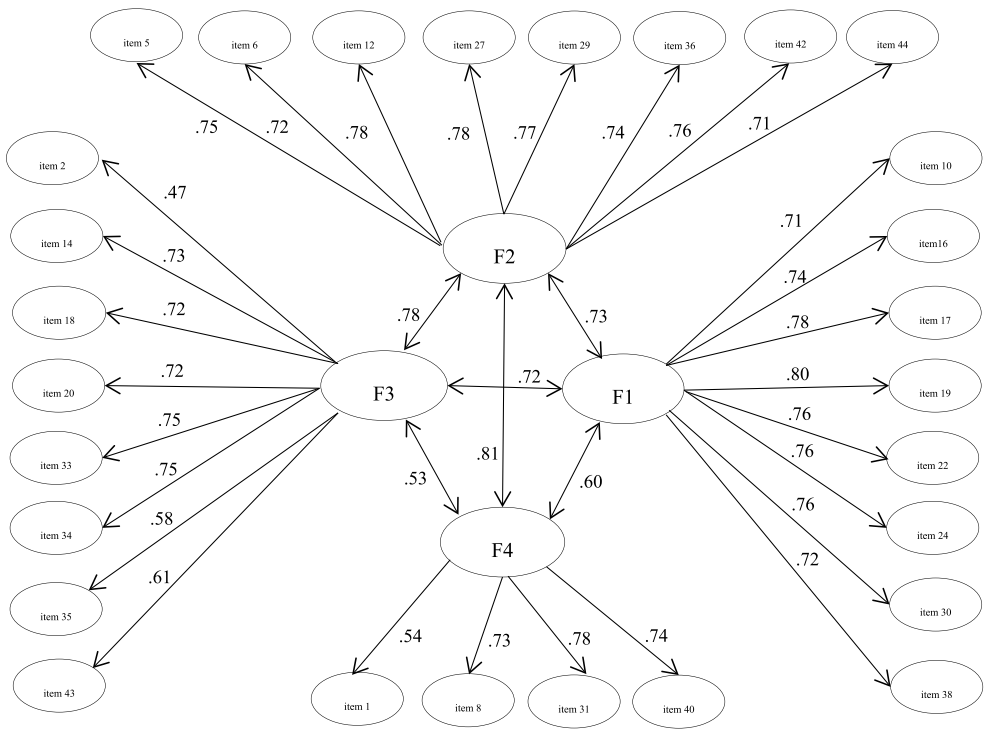
	Eigenvalue	Explained Variance (%)	Cumulative variance (%)	Proportion explained (%)	α	ωj
Intrapersonal motivation	7.90	18.37	18.37	33.80	0.93	0.86
Egoism/Hardness motivation	6.64	15.45	33.82	28.43	0.93	0.80
Interpersonal motivation	6.35	14.77	48.59	27.18	0.92	0.81
Malicious motivation	2.47	5.76	54.35	10.59	0.75	0.70

Total reliability α = 0.96 and ωj = 0.81.

TABLE 4 Correlations between CEMA-A factors ($N = 520$).

CEMA-A			
CEMA-A	Intrapersonal motivation	Egoism/Hardness motivation	Interpersonal motivation
Intrapersonal motivation	–		
Egoism/Hardness motivation	0.70***	–	
Interpersonal motivation	0.66***	0.62***	–
Malicious motivation	0.54***	0.56***	0.54***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.



Note: F1= Intrapersonal Motivation; F2= Egoism/Hardness Motivation; F3= Interpersonal Motivation; F4= Malicious Motivation

FIGURE 2
CEMA a four factor models. F1, Intrapersonal Motivation; F2, Egoism/Hardness Motivation; F3, Interpersonal Motivation; F4, Malicious motivation.

RMSEA = 0.06, CI = 0.058–0.063; SRMR = 0.048). Therefore, we can conclude that both the number of factors and the factor loading pattern of the items on the CEMA-A scale are similar for men and women.

10.3389/fpsyg.2023.1289209

Regarding metric invariance, the fit indices were acceptable according to gender ($\chi^2 = 2111.75$, df = 674, $p < 0.001$; CFI = 0.93; TLI = 0.92; NFI = 0.90; NNFI = 0.92; RMSEA = 0.06, CI = 0.057–0.062; SRMR = 0.051). The results show that the fit indices between the configural model and the metric model did not differ according to gender ($\Delta\chi^2 = 28.99$, Δdf = 24, $p = 0.220$) (see Table 5).

3.4 Differences in the sociodemographic data

To explore whether the CEMA-A questionnaire was useful for differentiating the motives for lying of people with different sociodemographic profiles, MANOVA was performed with the total sample (Sample 3). The Intrapersonal, Interpersonal, Egoism/Hardness, and Malicious Motivation scales were taken as dependent variables, and gender and educational level as independent variables. Significant differences were found according to gender [$F(1,1718) = 21.04$, $p < 0.001$]. Specifically, men scored higher than women in the Egoism/Hardness and Malicious Motivation scales (Table 6).

Regarding educational level, the MANOVA showed significant differences [$F(3,1719) = 1.9$, $p < 0.05$], particularly in the Interpersonal Motivation factor. However, after analyzing the *post-hoc* contrasts, no significant differences were found between the different levels of education (Table 7).

3.5 Convergent and discriminant validity

Convergent and discriminant validity was analyzed using Pearson's correlation between the CEMA-A and EPQ-R scales (Sample 1). All the CEMA-A factors correlated positively with Neuroticism and Psychoticism, and negatively with L scale, suggesting convergent validity (Table 8). The highest correlations

TABLE 5 Factor loading and internal consistency of latent variables.

Parameter estimate	Un-standard β	z	Standard B	Ω McDonald
Intrapersonal motivation → item 10	1		0.71	0.91
Intrapersonal motivation → item 16	1.08	24.48***	0.74	
Intrapersonal motivation → item 17	1.17	25.88***	0.78	
Intrapersonal motivation → item 19	1.20	26.14***	0.80	
Intrapersonal motivation → item 22	1.26	28.25***	0.76	
Intrapersonal motivation → item 24	1.14	24.62***	0.76	
Intrapersonal motivation → item 30	1.09	25.15***	0.76	
Intrapersonal motivation → item 38	1.13	24.16***	0.72	
Egoism/Hardness motivation → item 5	1		0.75	0.88
Egoism/Hardness motivation → item 6	0.99	27.98***	0.72	
Egoism/Hardness motivation → item 12	1.11	27.29***	0.78	
Egoism/Hardness motivation → item 27	1.08	27.43***	0.78	
Egoism/Hardness motivation → item 29	0.86	26.88***	0.77	
Egoism/Hardness motivation → item 36	1.01	25.52***	0.74	
Egoism/Hardness motivation → item 42	0.88	26.59***	0.76	
Egoism/Hardness motivation → item 44	0.78	23.04***	0.71	
Interpersonal motivation → item 2	1		0.47	0.84
Interpersonal motivation → item 14	1.46	15.5***	0.73	
Interpersonal motivation → item 18	1.47	15.57***	0.72	
Interpersonal motivation → item 20	1.43	15.58***	0.72	
Interpersonal motivation → item 33	1.56	15.72***	0.75	
Interpersonal motivation → item 34	1.48	15.81***	0.75	
Interpersonal motivation → item 35	1.22	16.28***	0.58	
Interpersonal motivation → item 44	1.21	15.80***	0.61	
Malicious motivation → item 1	1		0.54	0.79
Malicious motivation → item 8	1.42	17.10***	0.73	
Malicious motivation → item 31	1.58	17.61***	0.78	
Malicious motivation → item 40	1.26	17.24***	0.74	

***$p < 0.001$.

TABLE 6 Comparison of gender with CEMA-A Factors.

	Men (*N* = 760)		Women (*N* = 962)			
	M	*SD*	*M*	*SD*	*F*	η^2
Intrapersonal motivation	15.41	8.26	15.34	7.96	0.03	0.00
Interpersonal motivation	21.80	8.91	21.07	8.46	3	0.00
Egoism/Hardness motivation	16.20	8.65	13.67	7.19	43.93***	0.03
Malicious motivation	6.13	3.14	5.44	2.80	22.82***	0.01

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 7 Comparison of educational level with CEMA-A factors.

	Primary (*N* = 70)		Secondary (*N* = 240)		Baccalaureate (*N* = 736)		University (*N* = 677)			
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	η^2
Intrapersonal motivation	15.37	9.10	15.12	8.15	15.66	8.28	15.15	7.76	0.55	0.00
Interpersonal motivation	19.66	7.70	20.38	8.54	21.86	9.07	21.42	8.32	2.77*	0.00
Egoism/Hardness motivation	14.17	78.27	15.00	7.97	14.99	8.09	14.54	7.90	0.58	0.00
Malicious motivation	5.97	3.19	5.99	3.31	5.80	3.04	5.59	2.74	1.39	0.00

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 8 Correlations between CEMA-A factors and the EPQ-R personality questionnaire.

	EPQ-R			
CEMA-A	Extraversion	Neuroticism	Psychoticism	L scale
Intrapersonal motivation	−0.09*	0.37***	0.15***	−0.30***
Interpersonal motivation	−0.01	0.22***	0.11*	−0.31***
Egoism/Hardness motivation	0.05	0.21***	0.29***	−0.30***
Malicious motivation	−0.03	0.13***	0.31***	−0.18***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

were between Neuroticism and Intrapersonal Motivation ($r = 0.37$; $p < 0.001$), as well as between Psychoticism with Egoism/Hardness Motivation ($r = 0.29$; $p < 0.001$) and with Malicious Motivation ($r = 0.31$; $p < 0.001$). The Extraversion factor demonstrated discriminant validity, since no significant correlations were found with the CEMA-A factors, except for a low negative correlation with Intrapersonal Motivation ($r = -0.09$; $p < 0.05$).

3.6 Score stability (test–retest)

Vuong's (1989) test was applied to assess whether there were differences between the two non-nested models. Both models were verified as indistinguishable (variance test), and the fit of both models was equal for the focal population (non-nested likelihood ratio test) in the four categories of motivations to lie. The test–retest correlation was 0.86 for Intrapersonal Motivation (pretest: z 0.825, $p = 0.21$; retest: z 0.825, $p = 0.80$), 0.81 for Intrapersonal Motivation (pretest: z 1.248, $p = 0.10$; retest: z 1.258, $p = 0.90$), 0.93 for Egoism/Hardness Motivation (pretest: z 1.225, $p = 0.11$; retest: z 1.225, $p = 0.89$), and 0.77, for Malicious Motivation, (pretest: z 0.616, $p = 0.27$; retest: z 0.616, $p = 0.73$).

4 Discussion

The aim of this study was to verify the stability of the factorial structure of the CEMA-A questionnaire in the Spanish adult population. The results showed that the CEMA-A has adequate psychometric properties and is valid and reliable instrument to measure different motives behind every day lies. The new structure of the of the 28-item CEMA-A instrument was confirmed, through EFA and CFA, and the four-factor model containing the factors Intrapersonal Motivation, Interpersonal Motivation, Egoism/Hardness Motivation and Malicious Motivation, which concurs with the factorial structure of the preliminary study of 45 items (Armas-Vargas, 2021a). Moreover, the temporal stability of the measurement instrument scores was verified.

The general category Egoism–Hardness Motivation of the CEMA-A encompasses various subcategories of motives focused on obtaining personal benefits, such as instrumental motives (item 12 "to get what I want"; item 36 "to benefit from something"), motives related to manipulation of others (item 44 "because it is easier to manipulate others"; item 6 "to try to win in an argument with someone"), or motives related to showing a positive self-image (item 5 "to impress others"; item 27 "to give a good image of myself"). Instrumental and manipulative motives are related to those proposed in Levine et al.'s (2016) pancultural model: "non-monetary personal advantage," while the motives related to showing

a positive self-image of the CEMA-A are equivalent, in a way, to Levine et al.'s (2016) "self-image management." In the case of CEMA, it also includes the search for admiration. The general category Malicious Motivation of the CEMA-A includes content related to harming others (item 1 "to generate doubts about another person"; item 40 "to falsely accuse someone and cause harm") and has a certain similarity with the "malicious" category of Levine et al.'s (2016) pancultural model. The contents of the Egoism/Hardness and Malicious Motivations find a parallel with the type of serious lies proposed by DePaulo et al. (2004). According to these authors, people who tell serious lies try to profit from dubious deals, and lie instrumentally to get what they want, and to avoid punishment. The truths behind serious lies are often shameful, immoral, or illegal (DePaulo et al., 2004; Palena et al., 2021). Similarly, people high on Machiavellianism tend to engage in "immoral" behaviors to achieve their goals (Monaghan et al., 2020).

The general category Interpersonal Motivation focuses on motives that try to maintain positive social relationships and includes content on prosocial-empathy (item 35 "to make others feel good"; item 43 "to be kind and cordial with others"), sociability and agreeableness (Item 2 "to not offend others"), hide information that could cause harm (item 11 "to hide certain problems or difficulties"), or avoid problems with others (item 18 "to avoid problems with others"; item 34 "to hide something that I know is wrong"). The content of this category is related to the motives proposed in Levine et al.'s (2016) pancultural model, such as altruistic lies, social politeness, personal transgression, and evasion, respectively.

The general category Intrapersonal Motivation includes new content related to self-deception that has not been addressed in the area of motives for lying in the literature (Armas-Vargas, 2021a). Some of the reasons related to self-deception are "so as not to face the truth" (item 17), "for fear of facing reality" (item 19), "because it is difficult for me to accept things as they are" (item 30), where self-deception occurs through denial of a real problem and acting as if it did not exist (Goleman, 1985; Cohen, 2001; Zerubavel, 2006; Friedrichs, 2014). At some point in their lives, people may be exposed to unpleasant or traumatic situations that lead to the need for self-deception in order to survive the negative experience. Self-deception is the result of a functional and adaptive system in the protection of the self and the regulation of goals. It is not pathological in itself, since most people use it at some point in their lives (Sirvent et al., 2019). Some authors consider that self-deception can lead to a gain, such as improving self-image (Starek and Keating, 1991; Bachkirova, 2016). Other authors emphasize its function as an avoidance strategy, such as avoiding distress (Fingarette, 1969; Sackeim, 1983). It has also been proposed that self-deception may arise from selective attention, whereby certain information is ignored or dismissed, despite evidence (Greenwald, 1997; Sharot, 2011), Other research suggests that self-deception is a cognitive process of biasing information to obtain or maintain a false belief that may be beneficial or detrimental to oneself (Mei et al., 2022). A close relationship has been found between self-deception and deception of others (Lu and Chang, 2014). Self-deception functions as an automatic mechanism of protection and adaptation of the "I," which ultimately seeks to safeguard the psychic order (Armas-Vargas, 2020). These types of reasons fulfil the objective of hiding and/or denying evidence that we do not know or do not want to accept, which, if rejected, would leave us psychologically unprotected (Armas-Vargas, 2020, 2021a). Specifically, there is gain in self-deception: distress is avoided, real damage is minimized, and benefits such as subjective and interpersonal well-being and improving self-image are obtained (Friedrichs, 2014; Bachkirova, 2016). In the "process" of self-deception, many strategies that people use escape their control and awareness. Many implicit and automatic processes may be outside volitional reach (Bargh, 1990; Bargh et al., 2001). The evaluation of self-deception is therefore carried out as an experience already lived and past, whereby the person realizes (either by themselves or with the help of a professional) that they have been self-deceived (Armas-Vargas, 2017a,b, 2020). What is evaluated, therefore, is not the self-deception in the moment, but rather that the person was self-deceived (Martínez-Manrique, 2007).

In addition, intra-personally motivated lying includes personal and emotional reasons that evaluate content related to insecurity, problems of self-esteem, shame, or fear of what others will say (item 10 "because I do not accept myself as I am"; item 16 "so as not to reveal my own meanness"; item 22 "because I feel insecure"). These motives are responsible for adapting reality to our emotional and psychological needs, to protect our identity, self-esteem, and the image others have of us (Turner et al., 1975; Buller and Burgoon, 1996; Armas-Vargas, 2020, 2021a). Many of these emotional motives may be implicit or escape awareness (McClelland et al., 1989; Bargh and Chartrand, 1999; Bargh et al., 2001; Custers and Aarts, 2005).

The relationship between the CEMA-A and EPQ-R factors confirms convergent validity and evidences the role of personality in the motives for lying (Buller and Burgoon, 1996; Olson and Weber, 2004; McLeod and Genereux, 2008; Harhoff et al., 2023). One study found that coldness when lying (e.g., "I do not usually have remorse when I lie") was positively related to Psychoticism, whereas emotional self-regulation when lying (e.g., "I feel guilty when I'm caught in a lie") was negatively related. On the other hand, the Neuroticism factor has been found to positively correlate with Self-Deception, Insecurity, or Fear of Rejection and Criticism (Armas-Vargas, 2021b). Neuroticism has been related to the propensity to lie and to different types of lying (Phillips et al., 2011; Hart et al., 2020). Extraversion was not related to any of the CEMA-A factors, only showing a low negative correlation with Intrapersonal Motivation. Extraverted people tend to minimize, hide, and/or deny negative characteristics about themselves, to create a favorable impression to others (DePaulo et al., 1996; Tyler and Feldman, 2004; Armas-Vargas, 2021b).

Likewise, invariance analyses confirmed the equivalence for men and women of the measurements obtained by the instrument. Men scored higher in Egoism/Hardness Motivation and Malicious Motivation, which coincides with the pilot study (Armas-Vargas, 2021a). However, these differences must be taken with caution due to the small effect size found. However, Tyler and Feldman (2004) suggest that men and women may have different reasons for lying depending on circumstance. For women, lies are related to feigning positive feelings others, rather than being selfish (DePaulo et al., 1996; Tyler et al., 2006). A more self-centered lie may attempt to obtain a psychical rather than a monetary reward (DePaulo et al., 1996). These types of results can be explained through emotional variables, since women, tend to feel more distressed and see serious lies as less justifiable (DePaulo et al., 2004). Men tell more lies for their own benefit, despite potential harm to others, and more lies containing false information to manipulate others' impressions of them (Phillips et al., 2011).

The CEMA-A has shown adequate psychometric properties, although certain limitations should be considered. Firstly, there is no consensus around a single type of motive for lying (Seiter and Bruschke, 2007; Guthrie and Kunkel, 2013). Secondly, the four categories do not include all the reasons for lying, they are not exhaustive or exclusive. Although the CEMA-A was constructed by sampling the different motives for lying that appear in the literature, as well as collecting those such as self-deception that were not assessed through self-report, future research may find other reasons not identified thus far. Thirdly, response

biases may occur, both due to the content of the test itself (lies) and because it is a self-reported measure. This type of bias could be minimized by using a social desirability scale.

In future, analysis of the invariance in clinical and forensic samples, and in other cultures, could be interesting. Lying depends largely on the ethical and moral values of individuals and cultural conventions. Behaviors that are immoral in one culture may not be immoral in another (Kwiatkowska, 2015). Thus, it is important to identify whether the reasons for lying are similar, regardless of cross-cultural differences. Conversely, the reasons may vary, depending on whether the culture is individualistic or collectivist (Giles et al., 2019). In this sense, it could be of interest to adapt the CEMA-A to other cultures and verify its factorial invariance in different cultures. In addition, the CEMA-A questionnaire on motives for lying can be used to identify profiles of individuals according to their personality characteristics (e.g., the characteristics that define the person whose main motivation for lying is personal–emotional (fears, insecurity), as opposed to another whose motives are more focused on manipulating or instrumentalizing others). Previous research has shown that people with high anxiety, low self-esteem, and high Machiavellianism have motivations that will benefit them or others, whereas lies with protective motivation are associated with high empathy and low Machiavellianism (Cantarero et al., 2018). Furthermore, the CEMA-A could capture the motives for lying of different pathological populations, such as in the dark triad (psychopathy, Machiavellianism, and narcissism), where more malicious motives could appear. Michels et al. (2020) found a relationship between the dark triad and lying ability to achieve one's objectives, though this relationship was moderated by intelligence. In the same line, it could be of interest to use an instrument on lies in the forensic population, such as gender violence, or in contentious procedures for the custody of children. Intrapersonal motives may appear in victims of gender violence, while in aggressors the motivation would be more instrumental or malicious. In men convicted of gender violence, self-deception and an absolutist morality have been found to explain in some way the violent behavior against their partners (Vecina, 2018). Future studies could examine whether the CEMA-A questionnaire is useful for identifying populations that have a greater propensity to lie, depending on type of motive.

In summary, the CEMA-A questionnaire is based on an exhaustive review of the literature on motives for lying, including from social psychology models and personality psychology. The instrument therefore provides an empirical framework to identify the various motives for lying. They are grouped into four broad categories in which intrapersonal motivation related to self-deception and individual differences, previously little studied as motives for lying in the literature, play a major role. The CEMA-A has proven to be an adequate instrument for identifying categories, motives, situations, and moments that lead to lying; it is the first instrument in Spanish to assess motives for lying. These findings have important practical implications and could be a useful tool for analyzing the motives for lying in different clinical, forensic, and/or employment contexts. These types of lies may be interesting for future research on lying and understanding liars.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Research Ethics and Animal Welfare Committee of the University of La Laguna (Registration Number: CEIBA2023-3299). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

EA-V: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. RM: Project administration, Supervision, Validation, Visualization, Writing – review & editing. JH-C: Data curation, Formal analysis, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors thank the participants for their participation and involvement in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1289209/full#supplementary-material

References

Abad, F. J., Olea, J., Ponsoda, V., and y García, C. (2011). *Medición en ciencias sociales y de la salud [Measurement in Social and Educational Sciences]*. Madrid, España: Síntesis.

Aldridge, V. K., Dovey, T. M., and Wade, A. (2017). Assessing test-retest reliability of psychological measures: persistent methodological problems. *Eur. Psychol.* 22, 207–218. doi: 10.1027/1016-9040/a000298

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., and Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *J. Pers. Soc. Psychol.* 68, 804–825. doi: 10.1037/0022-3514.68.5.804

Arcimowicz, B., Cantarero, K., and y Soroko, E. (2015). Motivation and consequences of lying. A qualitative analysis of everyday lying. Forum qualitative. *Soc. Res.* 16:21. doi: 10.17169/fqs-16.3.2311

Ariely, D. (2012). *The (honest) truth about dishonesty*. New York: Harper Collins Publishers.

Armas-Vargas, E. (2017a). "Adaptación del cuestionario ATRAMIC: Personalidad y predisposición a mentir en adolescentes" in *Psicología Jurídica: Conocimiento y Práctica*. eds. C. Bringas and M. Novo (Sevilla: España)

Armas-Vargas, E. (2017b). "Autoengaño: Autoconocimiento y autoestima" in *En Actas del III Congreso Nacional de Psicología* (Oviedo, España: Consejo General de Psicología), 212–218.

Armas-Vargas, E. (2020). Autoengaño y mentira en adolescentes: Personalidad y autoestima. En A. M. Martín, F. Fariña and R. Arce (Eds.), *Psicología Jurídica y Forense: Investigación para la Práctica Profesional*. Madrid: España

Armas-Vargas, E. (2021a). Motivos para mentir (CEMA-A): predisposición a mentir y sesgo de respuesta. En L. Rodríguez Franco, D. Seijo and F. Fariña (Eds.), *Ciencia psicológica al servicio de la justicia y la ley*. Vigo: España.

Armas-Vargas, E. (2021b). La mentira como rasgo disposicional (Test ATRAMIC): personalidad y tendencia a mentir. En L. Rodríguez Franco, D. Seijo and F. Fariña (Eds.), *Ciencia psicológica al servicio de la justicia y la ley*. Vigo: España.

Bachkirova, T. (2016). A new perspective on self-deception for applied purposes. *New Ideas Psychol.* 43, 1–9. doi: 10.1016/j.newideapsych.2016.02.004

Bargh, J. A. (1990). "Auto-motives: preconscious determinants of social interaction" in *Handbook of motivation and cognition*. eds. E. T. Higgins and R. M. Sorrentino (New York: Guilford), 93–130.

Bargh, J. A., and Chartrand, T. L. (1999). The unbearable automaticity of being. *Am. Psychol.* 54, 462–479. doi: 10.1037/0003-066x.54.7.462

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., and Trötschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *J. Pers. Soc. Psychol.* 81, 1014–1027. doi: 10.1037/0022-3514.81.6.1014

Baumeister, R. F. (1993). "Lying to yourself: the enigma of self-deception" in *Lying and deception in everyday life*. eds. M. Lewis and C. Saarni (The Guilford Press)

Baumeister, R. F. (1997). *Evil: Inside human violence and cruelty*. New York: W.H. Freeman.

Bok, S. (1978). *Lying: Moral choice in public and private life*. New York, NY: Vantage Press.

Bond, C. F., and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personal. Soc. Psychol. Rev.* 10, 214–234. doi: 10.1207/s15327957pspr1003_2

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guildford Press.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (*2nd ed.*). New York, NY: Guilford Publications.

Browne, M. W., and Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivar. Behav. Res.* 24, 445–455. doi: 10.1207/s15327906mbr2404_4

Buller, D. B., and Burgoon, J. K. (1996). Interpersonal deception theory. *Commun. Theory* 6, 203–242. doi: 10.1002/9781118540190.wbeic170

Buller, D. B., Burgoon, J. K., White, C. H., and Ebesu, A. S. (1994). Interpersonal deception VII: behavioral profiles of falsification, equivocation, and concealment. *J. Lang. Soc. Psychol.* 13, 366–395. doi: 10.1177/0261927X94134002

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Cantarero, K., Van Tilburg, W. A. P., and Szarota, P. (2018). Differentiating everyday lies: a typology of lies based on beneficiary and motivation. *Personal. Individ. Differ.* 134, 252–260. doi: 10.1016/j.paid.2018.05.013

Cohen, S. (2001). *States of denial: Knowing about atrocities and suffering*. Cambridge, UK: Malden, Ma., Polity.

Curtis, D. A. (2021). You liar! Attributions of lying. *J. Lang. Soc. Psychol.* 40, 504–523. doi: 10.1177/0261927X21999692

Curtis, D. A., and Hart, C. L. (2015). Pinocchio's nose in therapy?: therapists' attitudes and beliefs toward client deception. *Int. J. Adv. Couns.* 37, 279–292. doi: 10.1007/s10447-015-9243-6

Curtis, D. A., and Hart, C. L. (2022). *Pathological lying: Theory, research, and practice*. Washington: American Psychological Association.

Custers, R., and Aarts, H. (2005). Positive affect as implicit motivator: on the nonconscious operation of behavioral goals. *J. Pers. Soc. Psychol.* 89, 129–142. doi: 10.1037/0022-3514.89.2.129

DePaulo, B. M., Ansfield, M. E., Kirkendol, S. E., and Boden, J. M. (2004). Serious lies. *Basic Appl. Soc. Psychol.* 26, 147–167. doi: 10.1207/s15324834basp2602&3_4

DePaulo, B. M., and Kashy, D. A. (1998). Everyday lies in close and casual relationships. *J. Pers. Soc. Psychol.* 74, 63–79. doi: 10.1037/0022-3514.74.1.63

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., and Epstein, J. A. (1996). Lying in everyday life. *J Pers. Soc. Psychol.* 70, 979–995. doi: 10.1037/0022-3514.70.5.979

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003b). Cues to deception. *Psychol. Bull.* 129, 74–118. doi: 10.1037/0033-2909.129.1.74

DePaulo, B. M., Wetzel, C., Sternglanz, R. W., and Walker Wilson, M. J. (2003a). Verbal and nonverbal dynamics of privacy, secrecy, and deceit. *J. Soc. Issues* 59, 391–410. doi: 10.1111/1540-4560.00070

Dunn, T. J., Baguley, T., and Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412. doi: 10.1111/bjop.12046

Ekman, P. (1985/2001). *Telling lies*. New York: W. W. Norton.

Ekman, P. (1989). "Why lies fail and what behaviors betray a lie" in *Credibility assessment*. ed. J. C. Yuille (Dordrecht, The Netherlands: Kluwer), 71–82.

Ennis, E., Vrij, A., and Chance, C. (2008). Individual differences and lying in everyday life. *J. Soc. Pers. Relat.* 25, 105–118. doi: 10.1177/0265407507086808

Erat, S., and Gneezy, U. (2012). White lies. *Manag. Sci.* 58, 723–733. doi: 10.1287/mnsc.1110.1449

Eysenck, H. J., and Eysenck, S. B. G. (1997). *EPQ-R. Cuestionario revisado de personalidad de Eysenck*. Madrid: TEA

Feldman, R. S., Forrest, J. A., and Happ, B. R. (2002). Self-presentation and verbal deception: do self-presenters lie more? *Basic Appl. Soc. Psychol.* 24, 163–170. doi: 10.1207/S15324834BASP2402_8

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

Fingarette, H. (1969). *Self-deception*. New York: Humanities Press.

Friedrichs, J. (2014). Useful lies: the twisted rationality of denial. *Philos. Psychol.* 27, 212–234. doi: 10.1080/09515089.2012.724354

Fullam, R. S., McKie, S., and Dolan, M. C. (2009). Psychopathic traits and deception: Functional magnetic resonance imaging study. *Br J Psychiatry.* 194, 229–235. doi: 10.1192/bjp.bp.108.053199

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Pract. Assess. Res. Eval.* 17, 1–13. doi: 10.7275/n560-j767

Gerlach, P., Teodorescu, K., and Hertwig, R. (2019). The truth about lies: a meta-analysis on dishonest behavior. *Psychol. Bull.* 145, 1–44. doi: 10.1037/bul0000174

Giammarco, E. A., Atkinson, B., Baughman, H. M., Veselka, L., and Vernon, P. A. (2013). The relation between antisocial personality and the perceived ability to deceive. *Personality and Individual Differences* 54, 246–250. doi: 10.1016/j.paid.2012.09.004

Giles, R. M., Rothermich, K., and Pell, M. D. (2019). Differences in the evaluation of prosocial lies: a cross-cultural study of Canadian, Chinese, and German adults. *Front. Commun.* 4:38. doi: 10.3389/fcomm.2019.00038

Gil-Escudero, G., and Martínez-Arias, M. R. (2001). "Metodología de encuestas [Survey methodology]" in *Métodos, Diseños y Técnicas de Investigación Psicológica*. ed. M. J. Navas (Universidad Nacional de Educación a Distancia)

Goleman, D. (1985). *Vital lies, simple truths: The psychology of self-deception*. New York, NY: Simon and Schuster.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *Am Psychol.* 52, 1182–1186. doi: 10.1037/0003-066X.52.11.1182

Gudjonsson, G. H., and Y Sigurdsson, J. F. (2004). The relationship of suggestibility and compliance with self-deception and other-deception. *Psychol Crime Law.* 10, 447–453. doi: 10.1080/10683160310001634278

Guthrie, J., and Kunkel, A. (2013). Tell me sweet (and not-so-sweet) little lies: deception in romantic relationships. *Commun. Stud.* 64, 141–157. doi: 10.1080/10510974.2012.755637

Harhoff, N., Reinhardt, N., Reinhard, M. A., and Mayer, M. (2023). Agentic and communal narcissism in predicting different types of lies in romantic relationships. *Front. Psychol.* 14:1146732. doi: 10.3389/fpsyg.2023.1146732

Hart, C. L., Jones, J. M., Terrizzi, J. A., and Curtis, D. A. (2019). Development of the lying in everyday situations scale. *Am. J. Psychol.* 132, 343–352. doi: 10.5406/amerjpsyc.132.3.0343

Hart, C., Lemon, R., Curtis, D., and Griffith, J. (2020). Personality traits associated with various forms of lying. *Psychol. Stud.* 65, 239–246. doi: 10.1007/s12646-020-00563-x

Haselton, M. G., Buss, D. M., Oubaid, V., and Angleitner, A. (2005). Sex, lies, and strategic interference: the psychology of deception between the sexes. *Personal. Soc. Psychol. Bull.* 31, 3–23. doi: 10.1177/0146167204271303

Hernández, J. A., and Betancort, M. (2018). ULLRtoolbox. Available at: https://sites.google.com/site/ullrtoolbox/

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi: 10.1007/BF02289447

Jensen, L., Arnett, J., Feldman, S., and Cauffman, E. (2004). The right to do wrong: lying to parents among adolescents and emerging adults. *J. Youth Adolesc.* 33, 101–112. doi: 10.1023/B:JOYO.0000013422.48100.5a

Kashy, D. A., and DePaulo, B. M. (1996). Who lies? *J. Pers. Soc. Psychol.* 70, 1037–1051. doi: 10.1037/0022-3514.70.5.1037

Keltner, D., and Buswell, B. N. (1996). Evidence for the distinctness of embarrassment, shame, and guilt: a study of recalled antecedents and facial expressions of emotion. *Cognit. Emot.* 10, 155–172. doi: 10.1080/026999396380312

Kline, R. B. (2011). "Convergence of structural equation modeling and multilevel modeling" in *The SAGE handbook of innovation in social research methods*. eds. M. Williams and W. P. Vogt (Thousand Oaks, CA: Sage), 562–589.

Knapp, M. L., and Comadena, M. A. (1979). Telling it like it isn't: a review of theory and research on deceptive communications. *Hum. Commun. Res.* 5, 270–285. doi: 10.1111/j.1468-2958.1979.tb00640.x

Kwiatkowska, A. (2015). "How do others deceive? Cultural aspects of lying and cheating" in *The small and big deceptions: In psychology and evolutionary sciences perspective*. eds. A. Kwiatkowska and A. Łukasik (Rzeszów, Poland: Wydawnictwo), 46–72.

Levine, T. R. (2014). Truth-default theory (TDT): a theory of human deception and deception detection. *J. Lang. Soc. Psychol.* 33, 378–392. doi: 10.1177/0261927X14535916

Levine, T. R., Ali, M. V., Dean, M., Abdulla, R. A., and Garcia-Ruano, K. (2016). Toward a pan-cultural typology of deception motives. *J. Intercult. Commun. Res.* 45, 1–12. doi: 10.1080/17475759.2015.1137079

Levine, T. R., Kim, R. K., and Hamel, L. M. (2010). People lie for a reason: three experiments documenting the principle of veracity. *Commun. Res. Rep.* 27, 271–285. doi: 10.1080/08824096.2010.496334

Levine, T. R., Serota, K. B., Carey, F., and Messer, D. (2013). Teenagers lie a lot: a further investigation into the prevalence of lying. *Commun. Res. Rep.* 30, 211–220. doi: 10.1080/08824096.2013.806254

Levine, T. R., Serota, K. B., and Punyanunt-Carter, N. M. (2022). Sender and receiver lie frequencies and motives: testing predictions from truth-default theory. *South Commun. J.* 87, 220–234. doi: 10.1080/1041794X.2022.2052745

Lindskold, S., and Walters, P. S. (1983). Categories for acceptability of lies. *J. Soc. Psychol.* 120, 129–136. doi: 10.1080/00224545.1983.9712018

Lu, H. J., and Chang, L. (2014). Deceiving yourself to better deceive high-status compared to equal-status others. *Evol. Psychol.* 12, 635–654. doi: 10.1177/147470491401200310

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* 1, 130–149. doi: 10.1037/1082-989X.1.2.130

Masip, J., Garrido, E., and y Herrero, C. (2004). Defining deception. *Anales de Psicología* 20, 147–171.

Martínez-Manrique, F. (2007). Attributions of Self-Deception. *Teorema* 26, 131–143.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* 45, 633–644. doi: 10.1509/jmkr.45.6.633

McArthur, J., Jarvis, R., Bourgeois, C., and Ternes, M. (2022). Lying motivations: exploring personality correlates of lying and motivations to lie. *Can. J. Behav. Sci.* 54, 335–340. doi: 10.1037/cbs0000328

McClelland, D. C., Koestner, R., and Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychol. Rev.* 96, 690–702. doi: 10.1037/0033-295X.96.4.690

McCornack, S. A., and Levine, T. R. (1990). When lies are uncovered: emotional and relational outcomes of discovered deception. *Commun. Monogr.* 57, 119–138. doi: 10.1080/03637759009376190

McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., and Zhu, X. (2014). Information manipulation theory 2: a propositional theory of deceptive discourse production. *J. Lang. Soc. Psychol.* 33, 348–377. doi: 10.1177/0261927X14534656

McDonald, R. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McLeod, B. A., and Genereux, R. L. (2008). Predicting the acceptability and likelihood of lying: the interaction of personality with type of lie. *Personal. Individ. Differ.* 45, 591–596. doi: 10.1016/j.paid.2008.06.015

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., and Ariely, D. (2009). Too tired to tell the truth: self-control resource depletion and dishonesty. *J. Exp. Soc. Psychol.* 45, 594–597. doi: 10.1016/j.jesp.2009.02.004

Mei, D., Ke, Z., Li, Z., Zhang, W., Gao, D., and Yin, L. (2022). Self-deception: distorted metacognitive process in ambiguous contexts. *Hum. Brain Mapp.* 44, 948–969. doi: 10.1002/hbm.26116

Metts, S. (1994). "Relational transgressions" in *The dark side of interpersonal communication*. eds. W. R. Cupach and B. H. Spitzberg (Hillsdale, NJ: Erlbaum)

Michels, M., Molz, G., and Bermpohl, F. M. (2020). The ability to lie and ist relations tot he dark triad and general intelligence. *Personal. Individ. Differ.* 166:110195. doi: 10.1016/j.paid.2020.110195

Millar, K. U., and Tesser, A. (1988). Deceptive behavior in social relationships: a consequence of violated expectations. *J. Psychol. Interdisc. Appl.* 122, 263–273. doi: 10.1080/00223980.1988.9915514

Miller, G. R., and Stiff, J. B. (1993). *Deceptive communication*. Newbury Park: Sage.

Monaghan, C., Bizumic, B., Williams, T., and Sellbom, M. (2020). Two-dimensional Machiavellianism: conceptualization, theory, and measurement of the views and tactics dimensions. *Psychol. Assess.* 32, 277–293. doi: 10.1037/pas0000784

Moshagen, M., Zettler, I., and Hilbig, B. E. (2020). Measuring the dark core of personality. *Psychol. Assess.* 32, 182–196. doi: 10.1037/pas0000778

Mosher, D. L. (1968). Measurement of guilt in females by self-report inventories. *J. Consult. Clin. Psychol.* 32, 690–695. doi: 10.1037/h0026589

Muñiz, J., and Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema* 31, 7–16. doi: 10.7334/psicothema2018.291

Muzinic, L., Kozaric-Kovacic, D., and Marinic, I. (2016). Psyquiatric aspectos of normal and pathological lying. *Int. J. Law Psychiatry* 46, 88–93.

Nyberg, D. (1993). *The varnished truth*. Chicago: University of Chicago Press.

Olson, K. R., and Weber, D. A. (2004). Relations between big five traits and fundamental motives. *Psychol. Rep.* 95, 795–802. doi: 10.2466/PR0.95.7.795-802

Palena, N., Caso, L., Cavagnis, L., and Greco, A. (2021). Profiling the interrogee: applying the person-centered approach in investigative interviewing research. *Front. Psychol.* 12:722893. doi: 10.3389/fpsyg.2021.722893

Phillips, M. C., Meek, S. W., and Vendemia, J. (2011). Understanding the underlying structure of deceptive behaviors. *Personal. Individ. Differ.* 50, 783–789. doi: 10.1016/j.paid.2010.12.031

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *J. Psychoeduc. Assess.* 29, 347–363. doi: 10.1177/0734282911406661

Sackeim, H. A. (1983). "Self-deception, self-esteem, and depression: The adaptive value of lying to oneself" in *Empirical studies of psychoanalytic theories*. ed. J. Masling (Hillsdale, NJ: Erlbaum), 101–157.

Schooler, J. W., and Schreiber, C. (2005). To know or not to know: Consciousness, metaconsciousness, and motivation. in *Social motivation: Conscious and non-conscious processes*. J. P. Forgas, K. R. Forgas and W. von Hippel (Eds.) (New York: Cambridge University Press), 351–372.

Seiter, J. S., and Bruschke, J. (2007). Deception and emotion: the effects of motivation, relationship type, and sex on expected feelings of guilt and shame following acts of deception in United States and Chinese samples. *Commun. Stud.* 58, 1–16. doi: 10.1080/10510970601168624

Seiter, J. S., Bruschke, J., and Bai, C. (2002). The acceptability of deception as a function of perceivers' culture, deceiver's intention, and deceiver-deceived relationship. *West. J. Commun.* 66, 158–180. doi: 10.1080/10570310209374731

Serota, K. B., and Levine, T. R. (2015). A few prolific liars: variation in the prevalence of lying. *J. Lang. Soc. Psychol.* 34, 138–157. doi: 10.1177/0261927X14528804

Serota, K. B., Levine, T. R., and Boster, F. J. (2010). The prevalence of lying in America: three studies of self-reported lies. *Hum. Commun. Res.* 36, 2–25. doi: 10.1111/j.1468-2958.2009.01366.x

Serota, K. B., Levine, T. R., and Docan-Morgan, T. (2022). Unpacking variation in lie prevalence: prolific liars, bad lie days, or both? *Commun. Monogr.* 89, 307–331. doi: 10.1080/03637751.2021.1985153

Sharot, T. (2011). The optimism bias. *Curr. Biol.* 21, R941–R945.

Sirvent, C., Herrero, J., de la Villa Moral, M., and Rodríguez, F. J. (2019). Evaluation of self-deception: Factorial structure, reliability and validity of the SDQ-12 (Self-Deception Questionnaire). *PLoS ONE* 14:e0210815. doi: 10.1371/journal.pone.0210815

Solomon, R. C. (2009). "Self, deception, and self-deception in philosophy" in *The philosophy of deception*. ed. C. Martin (New York: Oxford University Press), 15–36.

Starek, J. E., and Keating, C. F. (1991). Self-deception and its relationship to success in competition. *Basic Appl. Soc. Psych.* 12, 145–155. doi: 10.1207/s15324834basp1202_2

Tangney, J. P., Miller, R. S., Flicker, L., and Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *J. Pers. Soc. Psychol.* 70, 1256–1269. doi: 10.1037/0022-3514.70.6.1256

Teasdale, K., and Kent, G. (1995). The use of deception in nursing. *J. Med. Ethics* 21, 77–81. doi: 10.1136/jme.21.2.77

Timmerman, M. E., and Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychol. Methods* 16, 209–220. doi: 10.1037/a0023353

Touré-Tillery, M., and Fishbach, A. (2014). How to measure motivation: a guide for the experimental social psychologist. *Soc. Personal. Psychol. Compass* 8, 328–341. doi: 10.1111/spc3.12110

Trivers, R. (2002). "Self-deception in the service of deceit" in *Natural selection and social theory: Selected papers of Robert Trivers*. ed. R. Trivers (Oxford: Oxford University Press), 255–293.

Turner, R. E., Edgley, C., and Olmstead, G. (1975). Information control in conversations: honesty is not always the best policy. *Kansas J. Sociol.* 11, 69–89. doi: 10.17161/STR.1808.6098

Tyler, J. M., and Feldman, R. S. (2004). Truth, lies, and self-presentation: how gender and anticipated future interaction relate to deceptive behavior. *J. Appl. Soc. Psychol.* 34, 2602–2615. doi: 10.1111/j.1559-1816.2004.tb01994.x

Tyler, J. M., Feldman, R. S., and Reichert, A. (2006). The price of deceptive behavior: disliking and lying to people who lie to us. *J. Exp. Soc. Psychol.* 42, 69–77. doi: 10.1016/j.jesp.2005.02.003

Vecina, M. L. (2018). How can men convicted of violence against women feel moral while holding sexist and violent attitudes? A homeostatic moral model based on self-deception. *Am. J. Mens Health* 12, 1554–1562. doi: 10.1177/1557988318774218

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and its implications for professional practice*. Chichester: John Wiley and Sons.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. West Sussex, UK: John Wiley.

Vrij, A., and Ganis, G. (2014). "Theories in deception and lie detection" in *Credibility assessment: Scientific research and applications*. eds. D. C. Raskin, C. R. Honts and J. C. Kircher (Elsevier Academic Press)

Vuong, H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333. doi: 10.2307/1912557

Zerubavel, E. (2006). *The elephant in the room: Silence and denial in everyday life*. Oxford: Oxford University Press.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 12 January 2024
DOI 10.3389/fpsyg.2023.1326946

OPEN ACCESS

EDITED BY
Carmelo Mario Vicario,
University of Messina, Italy

REVIEWED BY
Julio C. Penagos-Corzo,
University of the Americas Puebla, Mexico
Zhen Zhang,
Henan Normal University, China

*CORRESPONDENCE
Alexander Korotkov
✉ korotkov@ihb.spb.ru

RECEIVED 24 October 2023
ACCEPTED 29 December 2023
PUBLISHED 12 January 2024

CITATION
Myznikov A, Korotkov A, Zheltyakova M, Kiselev V, Masharipov R, Bursov K, Yagmurov O, Votinov M, Cherednichenko D, Didur M and Kireev M (2024) Dark triad personality traits are associated with decreased grey matter volumes in 'social brain' structures.
Front. Psychol. 14:1326946.
doi: 10.3389/fpsyg.2023.1326946

Dark triad personality traits are associated with decreased grey matter volumes in 'social brain' structures

Artem Myznikov[1], Alexander Korotkov[1]*, Maya Zheltyakova[1], Vladimir Kiselev[1], Ruslan Masharipov[1], Kirill Bursov[1], Orazmurad Yagmurov[1], Mikhail Votinov[2], Denis Cherednichenko[1], Michael Didur[1] and Maxim Kireev[1,3]

[1]Russian Academy of Science, N.P. Bechtereva Institute of Human Brain, Saint Petersburg, Russia, [2]Department of Psychiatry, Psychotherapy and Psychosomatics, Medical Faculty, RWTH Aachen University, Aachen, Germany, [3]Saint Petersburg State University, Saint Petersburg, Russia

Introduction: Personality traits and the degree of their prominence determine various aspects of social interactions. Some of the most socially relevant traits constitute the Dark Triad – narcissism, psychopathy, and Machiavellianism – associated with antisocial behaviour, disregard for moral norms, and a tendency to manipulation. Sufficient data point at the existence of Dark Triad 'profiles' distinguished by trait prominence. Currently, neuroimaging studies have mainly concentrated on the neuroanatomy of individual dark traits, while the Dark Triad profile structure has been mostly overlooked.

Methods: We performed a clustering analysis of the Dirty Dozen Dark Triad questionnaire scores of 129 healthy subjects using the k-means method. The variance ratio criterion (VRC) was used to determine the optimal number of clusters for the current data. The two-sample t-test within the framework of voxel-based morphometry (VBM) was performed to test the hypothesised differences in grey matter volume (GMV) for the obtained groups.

Results: Clustering analysis revealed 2 groups of subjects, both with low-to-mid and mid-to-high levels of Dark Triad traits prominence. A further VBM analysis of these groups showed that a higher level of Dark Triad traits may manifest itself in decreased grey matter volumes in the areas related to emotional regulation (the dorsolateral prefrontal cortex, the cingulate cortex), as well as those included in the reward system (the ventral striatum, the orbitofrontal cortex).

Discussion: The obtained results shed light on the neurobiological basis underlying social interactions associated with the Dark Triad and its profiles.

KEYWORDS

machiavellianism, psychopathy, narcissism, K-means, vbm, emotional regulation, empathy, reward system

1 Introduction

The 'social brain' is one of the brain basis model of social interactions that incorporates many elements of various systems involved in socially significant aspects of human behavior (Adolphs, 2009). According to recent studies, the 'social brain' model comprises components of the reward system (Bhanji and Delgado, 2014), the Theory of Mind (ToM) neural system and its various domains, affective and cognitive, as well as the systems supporting empathy

(Maliske and Kanske, 2022). However, despite considerable research efforts directed at the regularities of the structural and functional organization of such systems, their neurobiological foundations remain uncertain.

Hence, a major goal of neurophysiological research should be to investigate how these brain systems are engaged in human social behavior. Such research on the neurobiological basis of human social behavior should take into account the influence of social intelligence or other personality characteristics (such as the Dark Triad) that might be associated with both the characteristics of human social activity and their neurobiological basis (Gundogdu et al., 2017; Segel-Karpas and Lachman, 2018; Back, 2021). We assume that such framework has already proved to be fruitful in studies that address the brain foundations of social behavior. For example, in recent morphometric and fMRI studies, we have found a link between gray matter morphometric characteristics of the basal ganglia (caudate nuclei) and its functional connectivity with the ToM neuronal system components, and social intelligence characteristics (Myznikov et al., 2021; Votinov et al., 2021). Using the Guilford-Sullivan test (O'Sullivan and Guilford, 1976), we reported higher gray matter volumes in the caudate nuclei of subjects with high social intelligence scores. Moreover, the level of social intelligence positively correlated with the degree of functional connectivity between the head of the right caudate nucleus and the brain areas associated with the ToM network, i.e., the right temporoparietal junction and the precuneus.

From the perspective of personality psychology, altogether with social intelligence, social behavior is affected by a set of non-pathological personality traits (non-pathological personalities), first described in theory of the Dark Triad by Paulhus and Williams (2002). The triad includes: (1) narcissism – characterized by excessive self-esteem and arrogance (Levy et al., 2011); (2) psychopathy – characterized by a lack of empathy and remorse, as well as such phenotypic domains as disinhibition, arrogance, and courage (Patrick, 2022); (3) Machiavellianism – characterized by a tendency to manipulate and use others for manipulator's purposes (Jones and Paulhus, 2009). According to a number of researchers, the combination of these traits constitutes the Dark Core of personality conceived as the Dark Factor of Personality, or the D-factor (Moshagen et al., 2018). Moshagen et al. (2018) defines the D-factor as "the tendency to maximize one's individual utility — disregarding, accepting, or malevolently provoking disutility for others —accompanied by beliefs that serve as justifications." Dark traits are considered as specific manifestations of a general, basic dispositional behavioral tendency that in fact manifests the Dark Factor of Personality. It is believed that individuals with more prominent dark personality traits tend to be more prone to antisocial and dangerous behavior, which cannot but affect the nature of social interactions (Lämmle and Ziegler, 2014; Sijtsema et al., 2019; Pechorro et al., 2022). Accordingly, the meta-analysis by Muris et al. (2017) demonstrated the connection between the dark traits and a number of unfavorable psychosocial factors. Namely, the level of narcissism was associated with difficulties in interpersonal relationships, while Machiavellianism – with difficulties in interpersonal relationships and antisocial tactics. The largest number of adverse psychosocial factors was associated with psychopathy and included aggression, socio-emotional deficit, sexual behavior disorders, and antisocial tactics.

Based on the literature, the sum of the test results (D-factor) is widely used to characterize dark personality traits. There are even well-founded recommendations to use composite scores rather than subscale scores for one of the dark triad tests because subscales contain small amounts of reliable variance beyond the general factor (Persson et al., 2019). At the same time, as some studies show, the Dark Triad construct is more complex than a mere sum of traits – rather a function of their multifaceted interaction, which poses the main problem of psychometric measurement concerning the Dark Triad (Kam and Zhou, 2016; Trahair et al., 2020; Truhan et al., 2021). The literature discusses whether the Dark Triad traits are independent behavioral predictors or whether they should be combined within the framework of an integral assessment of the complex of traits (the dark core of personality). In favor of this view, a recent study by McLarnon and colleagues used exploratory bifactor modeling by structural equations (B-ESEM) of the results of the SD3 questionnaire (McLarnon, 2022). This approach helped identify latent profiles of the triad in a healthy population based on four factors (narcissism, Machiavellianism, psychopathy, and the general D-factor). The profiles are defined as follows: (1) troublemakers (high levels of narcissism, psychopathy, and D-factor, a low level of Machiavellianism); (2) self-absorbed (low levels of Machiavellianism, psychopathy, and D-factor, a high level of narcissism); (3) manipulators (high levels of Machiavellianism and D-factor, low levels of psychopathy and narcissism); (4) exploiters (high levels of psychopathy and Machiavellianism, low levels of narcissism and D-factor). In another study, using the structural classification method, three categories were identified: benevolent (low machiavellianism, psychopathy and narcissism), intermediate malevolent (medium machiavellianism, psychopathy and narcissism) and high malevolent (high machiavellianism, psychopathy and narcissism; see Garcia and MacDonald, 2017). The multifaceted structure of the triad has also been explored in other studies, some of which use another comprehensive questionnaire – the Dirty Dozen Dark Triad (DDDT, Garcia and Rosenberg, 2016; Maneiro et al., 2020).

Despite the heterogeneity of the results obtained through the integrative approach that simultaneously considers all the subtests of the Dark Triad questionnaire, recent studies validate its efficiency. In this regard, we need to emphasize that we have not been able to find separate neurobiological studies that account for such integrity of the Dark Triad subtests indicators. As of now, the most widespread in the field are studies of the neuroanatomic organization of individual Dark Triad traits based on specialized psychometric questionnaires. Few studies have shown a positive correlation between the level of Machiavellianism (measured by a separate expanded specialized psychometric test MACH-IV) and the volume of some structures, such as subcortical nuclei, the prefrontal cortex, and the insula (Verbeke et al., 2011; Gong et al., 2023). Likewise, the meta-analysis by De Brito et al. demonstrated a gray matter volume decrease in the medial orbitofrontal and dorsolateral prefrontal cortex in psychopathy (De Brito et al., 2021). Additionally, a number of studies have shown a positive correlation between the level of psychopathy (assessed using PCL-R questionnaires) and the volume of subcortical nuclei (Korponay et al., 2017; Lam et al., 2017), in particular, the putamen, caudate nuclei, although an inverse relationship is also reported (Vieira et al., 2015). Meanwhile, the level of narcissism assessed using an Neuropsychiatric Inventory questionnaire (NPI) positively correlated with the volume of a number of structures, including the medial and the ventromedial prefrontal cortex, the dorsolateral prefrontal and orbitofrontal cortex, the middle anterior cingulate cortex, and the insula (Nenadić et al., 2021). Notably, in the case of the pathological variant – the narcissistic personality disorder – the

degree of manifestation of the disease negatively correlated with the volume of the medial prefrontal and dorsolateral cortex (Nenadic et al., 2015). Summarizing the above studies, we should witness a clear lack of research on the neurobiological foundations underlying the general construct of the Dark Triad rather than its individual features.

As regards the neuroimaging studies that applied voxel morphometry, we would like to draw attention to at least two major weaknesses that we attempt to address in the present study. First, as mentioned, these cited studies consider individual Dark Triad traits within the framework of specialized tests overlooking the general construct (an integrative evaluation of all traits). Second, these studies tend to disregard various Dark Triad profiles. In view of the above, our work was designed to conduct a morphometric study of dark personality traits, taking heed of the integral characteristics of the Dark Triad. To this end, with the present study we tried to overcome the main, in our opinion, shortcomings of the extant literature by using the general Dark Triad questionnaire, DDDT, and the k-means data clustering algorithm that allows an integral assessment of the Dark Triad traits. This would allow us to adequately approach the multifactorial structure of the Dark Triad and identify groups based on the data-driven approach. Further, a morphometric analysis was performed for the obtained groups to assess the differences in gray matter volumes between different Dark Triad profiles. We did not formulate specific hypotheses about the particular localization of possible changes in the gray matter depending on the Dark Triad indicators, but expected to detect these changes in brain structures that are associated with the 'social brain', primarily with the ToM network.

2 Materials and methods

2.1 Participants

A total of 129 healthy right-handed volunteers (70 women and 59 men) participated in the study. All participants were 24.4 ± 4 years old, with no history of neurological or psychological disorders and no contraindications for magnetic resonance imaging. All subjects provided written informed consent prior to the study. All procedures were conducted in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the N.P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences.

2.2 Psychological testing and clustering procedure

The Russian version of the Dark Triad Dirty Dozen (Kornilova et al., 2015) is a 12-item self-report questionnaire of the three Dark Triad traits with the 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree) and with statements such as: "I tend to manipulate others to get my way" (Machiavellianism), "I tend to lack remorse" (psychopathy), and "I tend to want others to admire me" (narcissism). Scores for all three traits were summed and then transformed into z-scores for using in clustering procedure.

To account for the multifactorial structure of the Dark Triad and to reveal subgroups in our sample, we applied a clustering procedure using the k-means algorithm implemented in MATLAB. This method uses the squared Euclidean distance metric and the k-means++ algorithm for cluster center initialization (Arthur and Vassilvitskii, 2007). We did not use the latent profile analysis (LPA) based on structural equation modeling (SEM) results as previously described (McLarnon, 2022) primary due to the small sample size. Instead, the k-means algorithm is a simple and intuitive approach that is effective in both small (not below 50 subjects) and large data samples (Henry et al., 2015). The variance ratio criterion (VRC) was used to determine the optimal number of clusters for the current data. The maximum VRC value was obtained for $k = 2$.

2.3 Data acquisition and preprocessing

Magnetic resonance imaging was performed using a 3T Philips Achieva (Philips Medical Systems, Best, The Netherlands). Structural images were acquired using a T1-weighted pulse sequence (T1W-3D-FFE; repetition time [TR] = 2.5 ms; TE = 3.1 ms; 30° flip angle), recording 130 axial slices (field of view [FOV] = 240 × 240 mm; 256 × 256 scan matrix) of 0.94 mm thickness. All MRI scans were inspected for image artifacts and incidental brain abnormalities. All subjects were included in the study.

2.4 Voxel based morphometry analysis (VBM analysis)

The VBM analysis of structural data was performed with Statistical Parametric Mapping software (SPM12, Wellcome Department of Imaging Neuroscience, London, UK[1]) and the Computational Anatomy Toolbox 12 (CAT12[2]) running in MATLAB (MathWorks, Natick, MA). All structural data were manually reoriented to place their native-space origin at the anterior commissure. Images were corrected for magnetic field inhomogeneities and segmented into grey matter, white matter, and cerebrospinal fluid. Normalization to MNI space using the DARTEL (Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra) algorithm to a 1.5 mm isotropic adult template provided by the CAT12 toolbox was performed for segmented grey matter data. Finally, the grey matter segments were smoothed with a Gaussian smoothing kernel of 8 mm. The CAT12 toolbox provides an automated quality check protocol. Therefore, quality check control for all structural data was performed to obtain so-called image quality rating (IQR) scores that were later used as an additional covariate in the statistical analysis. In addition, total intracranial volumes (TIVs) were calculated to be used as a covariate.

2.5 Statistical analysis of VBM data

The statistical analysis was performed for two groups of subjects. For the VBM analysis, we included the following confounders (as covariates), which can affect VBM results: sex (male/female), age, TIV (using ANCOVA), and IQR scores. The two-sample t-test was performed to test the hypothesized differences in the gray matter volume (GMV). The threshold-free cluster enhancement (TFCE) implemented in the TFCE-toolbox[3] with 5000 permutations per test

1 www.fil.ion.ucl.ac.uk/spm
2 http://dbm.neuro.uni-jena.de/cat.html
3 http://dbm.neuro.uni-jena.de/tfce

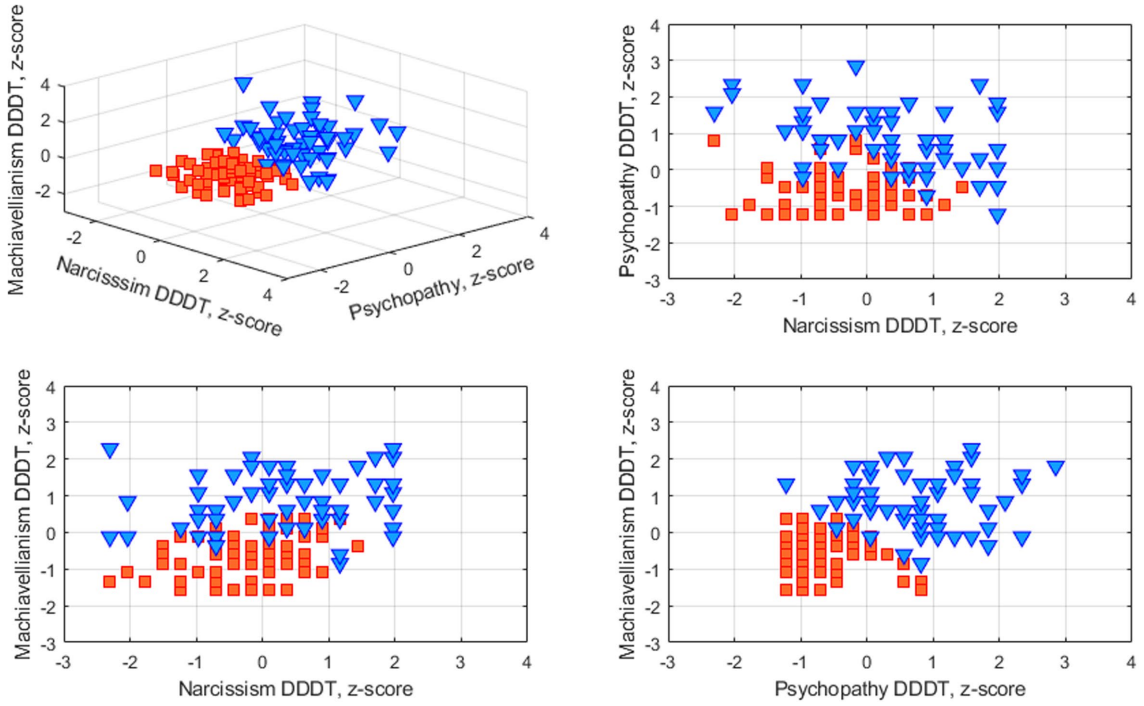


FIGURE 1
Results of DDDT data clustering using k-means algorithm.

was applied. Statistical parametric maps were created based on TFCE with $p < 0.001$ on the voxel level with the family-wise error (FWE) correction for multiple comparisons. The SPM results were visualized using the MRIcron toolbox.[4]

3 Results

3.1 DDDT data clustering

The results of data clustering are shown in Figure 1. Acquired clusters differed significantly in levels of DDDT subscales as well as in overall DDDT scores (see Table 1 and Figure 2). Clusters did not differ significantly in gender distribution (*chi-square* = 3.09, $p = 0.08$) or age ($p = 0.26$). Clusters were designated as "mid-to-high DT level" (57 subjects) and "low-to-mid DT level" (72 subjects).

3.2 GMV differences between low-to-mid and mid-to high DT levels

The VBM analysis revealed the GMV decrease in the mid-to-high DT level group in several regions including the prefrontal cortex (both medial and lateral, orbitofrontal cortex), the basal ganglia (the bilateral nucleus accumbens and putamen, the left caudate), the middle cingulate cortex as well as the right postcentral gyrus (see Table 2 and Figure 3; see also Supplementary material). Inverse contrast (High>Low) did not reveal any significant clusters.

4 Discussion

With this study, we aimed at assessing the impact Dark Triad profiles have on anatomical features of the brain. To define the profiles, we used a short DDDT questionnaire – a psychological tool for rapid integral assessment of the Dark Triad, consisting of 12 questions, with four questions per each of the Dark Triad traits – narcissism, psychopathy, Machiavellianism. All of these dark personality traits are complex constructs, and for each of these traits individual questionnaires determine a multifactorial structure (MACH-IV, NPI, SRP). Thus, according to the MACH-IV questionnaire, the following factors can be identified in the structure of Macchiavelinism: positive and negative interpersonal tactics, a cynical view of human nature, and disregard for moral norms (Muris et al., 2017). The multifactorial structure of narcissism includes lust for power, a sense of superiority, exhibitionism, feeling entitled, vanity, tendencies to exploit other people's resources, and self-confidence (del Rosario and White, 2005). Finally, the structure of psychopathy according to the SRP questionnaire includes interpersonal manipulation, callousness, erratic lifestyle, and a tendency toward criminal behavior (Massa and Eckhardt, 2017). A comparison of the DDDT questionnaire and individual tests for assessing dark personality traits showed that the DDDT allows evaluating only three factors of narcissism out of seven (entitlement, superiority, exhibitionism), two factors of Machiavellianism (manipulative tactics, disregard for moral norms) and one factor of psychopathy (callousness) (Muris et al., 2017).

4 https://www.nitrc.org/projects/mricron

TABLE 1 Demographics and scores (mean ± standard deviation) for different psychological scales in the revealed data clusters.

Psychological scale	Cluster 1 (Low-to-Mid DT Level)	Cluster 2 (Mid-to-High DT Level)	*p*-value
Age and gender			
Age	24.1 ± 3.8	24.9 ± 4.3	0.258
Gender (male/female)	28/44	31/26	0.079
Dirty dozen dark triade			
Narcissism	11.7 ± 2.9	13.8 ± 4.3	0.001
Psychopathy	6.3 ± 2.0	12.0 ± 3.4	<0.001
Machiavellianism	7.7 ± 2.3	14.0 ± 3.1	<0.001
General score	25.7 ± 4.7	39.8 ± 5.7	<0.001

DT, Dark Triad.

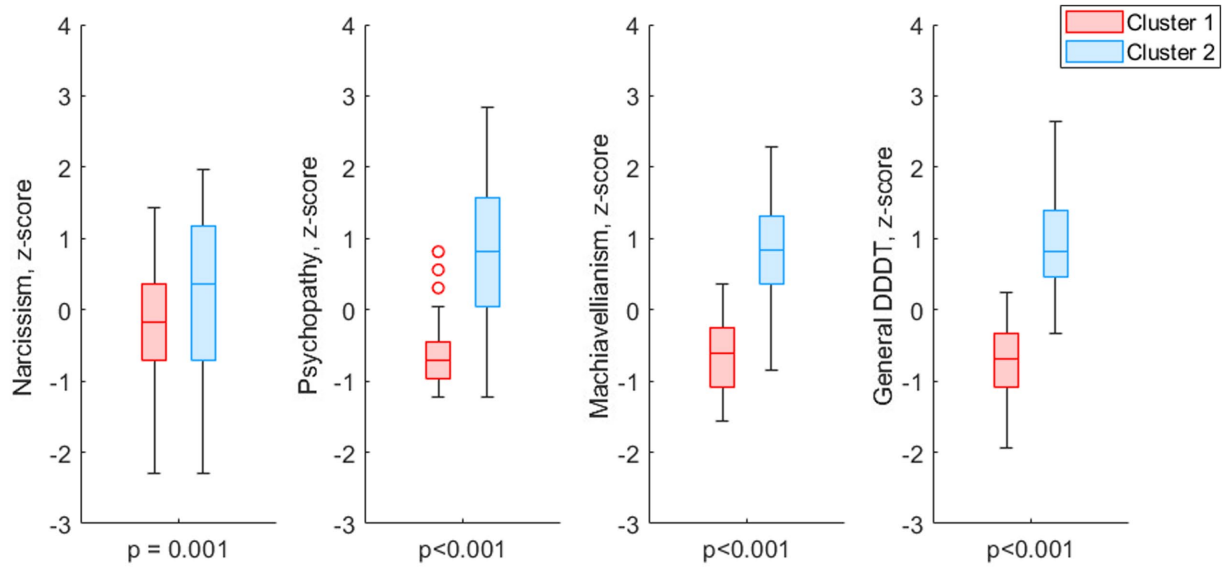


FIGURE 2
Boxplot of z-scores of DDDT subscales and general DDDT score for data clusters. Outliers are shown as red dots.

We find it important for understanding the psychological differences identified for the clustering groups.

While the vast majority of research articles on neuroanatomic correlates of Dark Triad were focused only on narcissism, psychopathy and machiavellianism separately, our data-driven clustering approach provided for a more integrative assessment. To isolate data-driven profiles using psychometric indicators of the Dark Dozen questionnaire, an algorithm for clustering data using k-means was applied. As a result, data were divided into two groups based on the clustering effectiveness evaluation. The further analysis showed that the obtained groups significantly differed in terms of each of the Dark Triad subscales, as well as its overall cumulative score according to the DDDT questionnaire. At the same time, the groups did not differ significantly in gender distribution and age. Hence, these results demonstrate the effectiveness of the chosen clustering method (k-means) in respect to our data. These groups were used in the further morphometric analysis that revealed a gray matter volume decrease in individuals with prominent dark personality traits in a number of structures, including the medial and dorsolateral prefrontal, the orbitofrontal cortex, subcortical nuclei (the nucleus accumbens and left shell), the middle cingulate cortex, and the right precentral gyrus.

The results of the morphometric analysis confirmed our assumption about differences between dark trait prominence levels manifested in the volume of structures related to the processing of socially significant information. We detected a cluster in the area of the medial prefrontal cortex that overlaps the ToM system area. One of the possible ways to describe the ToM system is to define cognitive and affective domains. The affective domain is usually related to understanding of the emotional states of others, while the cognitive domain is assumed to be involved in understanding thoughts and intentions (Molenberghs et al., 2016). We registered a decrease in gray matter volumes only for the medial prefrontal cortex (mPFC), an element of the affective domain. The involvement of the mPFC in the affective ToM processes has been shown in many studies (see for review Lieberman et al., 2019). More specifically, patients with damage to the ventral mPFC regions performed worse on the task of recognizing mental states than control groups – the task that should

10.3389/fpsyg.2023.1326946

TABLE 2 Results of VBM analysis (Low DT > High DT), TFCE voxel-level pFWE<0.001, k > 80.

No	Brain area	*k*	T (TFCE)-value	MNI coordinates		
				x	y	z
1	R MFG R SFG	1557	2744.2556	37.5	33	43.5
			2056.4529	30	10.5	57
2	R/L Accumbens R/L Putamen L Caudate R/L Septal Area	4457	2542.4211	−28.5	4.5	−7.5
			2279.8284	−7.5	3	−10.5
			2235.2615	12	21	−9
3	L MFG L SFG	1639	2497.1233	−28.5	61.5	−1.5
4	R SFG (medial)	289	2074.8376	1.5	64.5	22.5
5	R Postcentral G	121	2060.4263	15	−51	75
6	L MFG (medial)	146	2026.1814	−1.5	51	40.5
7	R MCC/PCC	102	2024.5991	10.5	−9	49.5
8	L Anterior OFC	97	2009.1361	−33	48	−18

MFG, middle frontal gyrus; SFG, superior frontal gyrus; MCC/PCC, middle/posterior cingulate cortex.

involve the affective ToM (Shamay-Tsoory et al., 2006; Shamay-Tsoory and Aharon-Peretz, 2007; Leopold et al., 2012). A decrease in the medial prefrontal cortex gray matter volume was associated with high levels of psychopathy (de Oliveira-Souza et al., 2008; Yang and Raine, 2009; Yang et al., 2010; Lam et al., 2017), although in some studies the dependence was reversed (Korponay et al., 2017).

A gray matter volume decrease in individuals with higher DT prominence was also detected in the right dorsolateral prefrontal cortex (dlPFC), one of the central executive network nodes that is also involved in processing of emotions and emotional stimuli. In particular, according to the term-based meta-analysis of Neurosynth, one of the terms associated with the cluster obtained in our study is reappraisal associated with the emotion regulation. Cognitive reassessment is one of the regulatory mechanisms during which a person re-evaluates a situation and its significance in order to change its emotional impact (Gross, 2015). It has been shown that psychopathy is characterized by low reliance on this mechanism, while for Machiavellianism and narcissism, this dependence has not been previously identified (Walker et al., 2022). The ventromedial and the dorsolateral prefrontal cortex, as previously shown, play different roles in emotion processing: for example, the dlPFC is involved in the control and regulation of emotional experience valence, while the ventromedial prefrontal cortex (vmPFC) may be involved in the suppression of arousal caused by emotional stimuli (Nejati et al., 2021). In addition, structural post-traumatic changes in the left dlPFC were associated with a high level of Machiavellianism according to the TDM-IV scale, in particular, with a high level on the "Machiavellian Views" scale (Cohen-Zimerman et al., 2017).

In addition, our study revealed a gray matter volume decrease in the left orbitofrontal cortex (OFC). There is evidence of the OFC involvement in processes associated with both affective and cognitive empathy (Brink et al., 2011). It is believed that the OFC may be involved in the modulation of empathy by various factors, such as social distance or gender (Hillis, 2014). In addition, the OFC is reportedly involved in emotional processing: a decrease in the OFC gray matter volume negatively correlated with the degree of emotional dysregulation (Petrovic et al., 2016). An OFC gray matter volume decrease was also associated with decreased ability to track dynamically changing emotions (Goodkind et al., 2012). Finally, according to a recent meta-analysis of morphometric studies, the gray matter volume in the orbitofrontal and prefrontal cortex negatively correlated with the degree of impulsivity (Pan et al., 2021) inherent in dark personality traits – psychopathy and to a lesser extent narcissism (Jones and Paulhus, 2011; Ball et al., 2018; Malesza and Kalinowski, 2021).

Summarizing all the above, the results of our study demonstrate a possible relationship between the prominence of the Dark Triad traits and the volume of structures associated with socio-emotional functions, such as empathy and emotional regulation. The neuroanatomic data agree with the outcomes of psychological works. Namely, a number of studies and meta-analyses have demonstrated a negative correlation between various components of emotional intelligence and the Dark Triad traits, in particular, Machiavellianism and psychopathy, although no correlation has been shown for narcissism (Miao et al., 2019; Michels and Schulze, 2021). At the same time, we observed no changes in the gray matter volume in structures related to sociocognitive functions, for example, the cognitive ToM domain. Manipulative behavior characterizes the Dark Triad and manifests itself in anatomical and functional features of the cognitive ToM domain structures.

Applying the morphometric analysis, we revealed an extensive cluster of gray matter volume reduction in individuals with a higher DT level in the structures of the mesolimbic system – the nucleus accumbens, the ventral striatum, the septal region – often referred to collectively as the Basal Forebrain (BF). These structures are anatomical correlates of reward processing and behavior regulation. Moreover, Hoffman and O'Connell attribute the above-mentioned areas to the social decision-making network in mammals (O'Connell and Hofmann, 2011). Additionally, Morelli and colleagues link activations in the septal region with variants of empathy (to pain, anxiety, happiness level) that predicted daily help to other people (Morelli et al., 2014). Likewise, numerous works indicate the BF role in altruistic behavior and charitable donations (Moll et al., 2006; Harbaugh et al., 2007; Kirk et al., 2016) and expectation of reward

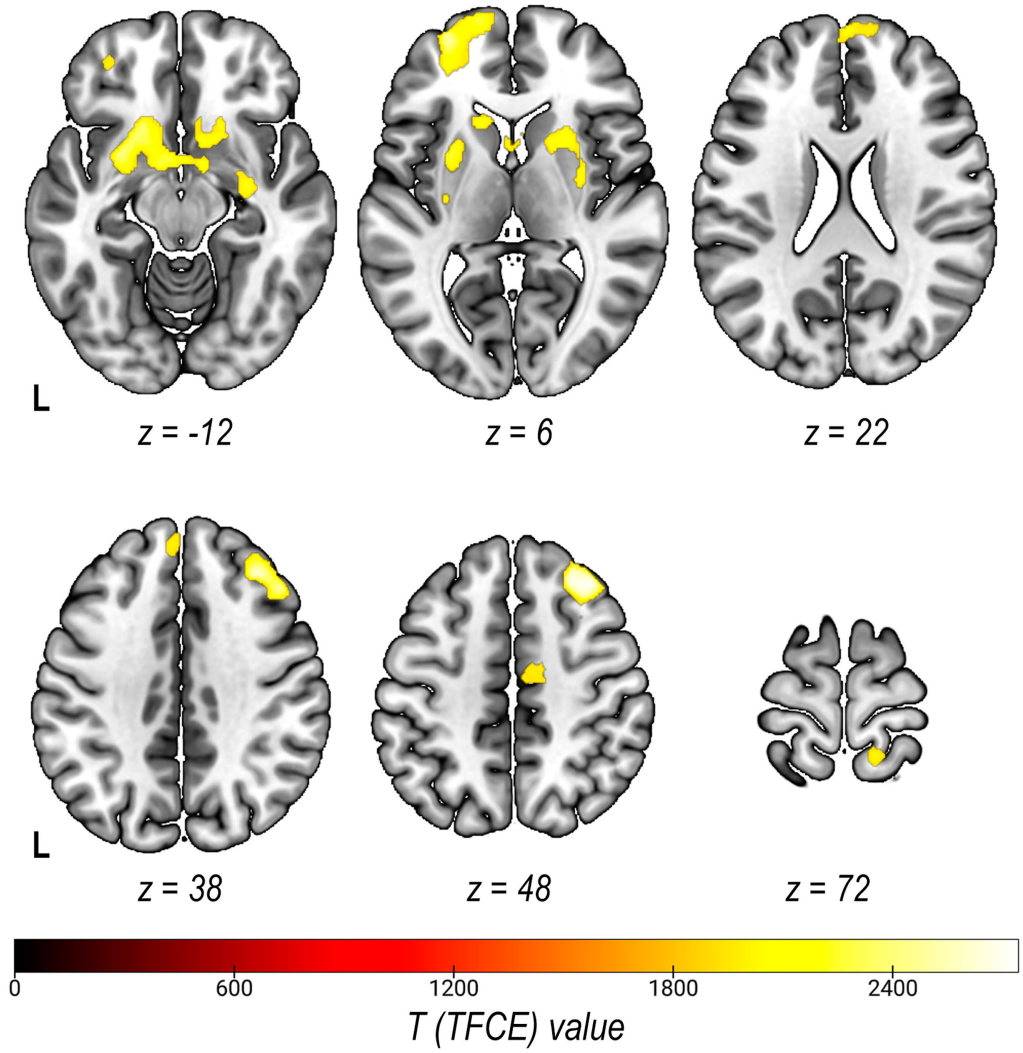


FIGURE 3
Statistical parametric maps of grey matter volume differences in subjects with Low and High DT Levels (contrast Low>High) at the TFCE voxel-level *pFWE* < 0.001, k > 80.

(Schultz et al., 1997). Many pathological conditions in which violations of various aspects of social behavior are observed (ASD, FTLD, etc.) are accompanied by structural and functional changes in the BF (Nickl-Jockschat et al., 2011; Riva et al., 2011; Convery et al., 2020; Schulz et al., 2023). Since a violation of social interactions is also observed in individuals with a higher DT level (see Introduction), the neuroanatomic basis of such antisocial behavior may be a lower volume of BF structures.

The present study has a limitation that is related to the psychological assessment technique we used for the Dark Triad. The DDDT, as mentioned above, while being a widely used simple and reliable instrument, does not fully cover the multifactorial structure of the individual traits of the Dark Triad as well as does not count associations between prominence of Dark Triad traits and both empathy (Pajevic et al., 2018; Heym et al., 2019) and aggression (Pailing et al., 2014; Jones and Neria, 2015). Potential solutions to this limitation for future research are (1) the use of the SD3 questionnaire, which is more inclusive of the multifactor structure of the Dark Triad, (2) the use of separate questionnaires for each of the Dark traits and (3) the use of separate questionnaires for associated traits like aggression and empathy. In addition, the use of structural equation modeling of psychological data on larger sample sizes would result in additional insight about the structure of the Dark Triad profile with the identification of their neuroanatomical correlates.

5 Conclusion

One way toward clarifying the nature of the 'social brain' is to describe the complex relations between psychometric indicators of dark personality traits. Our study helped identify neuroanatomic correlates of Dark Triad trait prominence levels via clustering data from the DDDT questionnaire. We were able to elucidate the neurobiological basis of social behavior in individuals with a higher DT level by analyzing gray matter volume variance in brain areas that provide for different aspects of social interactions. The volume decrease in structures associated with emotional regulation (OFC, vmPFC/dlPFC), empathy (OFC, MCC), and the reward system (basal

forebrain) complements assumptions about changes in the operation of these systems in individuals with various Dark Triad trait prominence levels.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Ethics Committee of the N.P. Bechtereva Institute of the Human Brain, St. Petersburg, Russia. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AM: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. AK: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. MZ: Formal analysis, Investigation, Writing – original draft. VK: Formal analysis, Visualization, Writing – original draft. RM: Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. KB: Formal analysis, Writing – review & editing. OY: Data curation, Methodology, Writing – review & editing. MV: Data curation, Investigation, Validation, Writing – review & editing. DC: Funding acquisition, Project administration, Resources, Writing – review & editing. MD: Funding acquisition, Resources, Supervision, Writing – review & editing. MK: Conceptualization, Data curation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was performed within the state assignment of the Ministry of Education and Science of Russian Federation (theme number AAAA-A19-122041500046-5).

Acknowledgments

We thank Diana Chopchik for her help in providing critical comments on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1326946/full#supplementary-material

References

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annu. Rev. Psychol.* 60, 693–716. doi: 10.1146/annurev.psych.60.110707.163514

Arthur, D., and Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, New Orleans, LA, United States.

Back, M. D. (2021). "Social interaction processes and personality" in *The handbook of personality dynamics and processes*. ed. J. F. Rauthmann (Amsterdam: Academic Press), 183–226.

Ball, L., Tully, R., and Egan, V. (2018). The influence of impulsivity and the dark triad on self-reported aggressive driving behaviours. *Accid. Anal. Prev.* 120, 130–138. doi: 10.1016/j.aap.2018.08.010

Bhanji, J. P., and Delgado, M. R. (2014). The social brain and reward: social information processing in the human striatum. *Wiley Interdiscip. Rev. Cogn. Sci.* 5, 61–73. doi: 10.1002/wcs.1266

Brink, T. T., Urton, K., Held, D., Kirilina, E., Hofmann, M., Klann-Delius, G., et al. (2011). The role of orbitofrontal cortex in processing empathy stories in 4- to 8-year-old children. *Front. Psychol.* 2:80. doi: 10.3389/fpsyg.2011.00080

Cohen-Zimerman, S., Chau, A., Krueger, F., Gordon, B., and Grafman, J. (2017). Machiavellian tendencies increase following damage to the left dorsolateral prefrontal cortex. *Neuropsychologia* 107, 68–75. doi: 10.1016/j.neuropsychologia.2017.11.007

Convery, R. S., Neason, M. R., Cash, D. M., Cardoso, M. J., Modat, M., Ourselin, S., et al. (2020). Basal forebrain atrophy in frontotemporal dementia. *Neuroimage Clin.* 26:102210. doi: 10.1016/j.nicl.2020.102210

De Brito, S. A., McDonald, D., Camilleri, J. A., and Rogers, J. C. (2021). Cortical and subcortical gray matter volume in psychopathy: a voxel-wise meta-analysis. *J. Abnorm. Psychol.* 130, 627–640. doi: 10.1037/abn0000698

de Oliveira-Souza, R., Hare, R. D., Bramati, I. E., Garrido, G. J., Azevedo Ignácio, F., Tovar-Moll, F., et al. (2008). Psychopathy as a disorder of the moral brain: fronto-temporo-limbic grey matter reductions demonstrated by voxel-based morphometry. *Neuroimage* 40, 1202–1213. doi: 10.1016/j.neuroimage.2007.12.054

del Rosario, P. M., and White, R. M. (2005). The narcisssistic personality inventory: test–retest stability and internal consistency. *Personal. Individ. Differ.* 39, 1075–1081. doi: 10.1016/j.paid.2005.08.001

Garcia, D., and MacDonald, S. (2017). Dark personality profiles: estimating the cluster structure of the dark triad. *Psych. J.* 6, 239–240. doi: 10.1002/pchj.175

Garcia, D., and Rosenberg, P. (2016). The dark cube: dark and light character profiles. *PeerJ* 4:e1675. doi: 10.7717/peerj.1675

Gong, X., Quan, F., Wang, L., Zhu, W., Lin, D., and Xia, L.-X. (2023). The relationship among regional gray matter volume in the brain, Machiavellianism and social aggression in emerging adulthood: a voxel-based morphometric study. *Curr. Psychol.* 42, 25160–25170. doi: 10.1007/s12144-022-03574-1

Goodkind, M. S., Sollberger, M., Gyurak, A., Rosen, H. J., Rankin, K. P., Miller, B., et al. (2012). Tracking emotional valence: the role of the orbitofrontal cortex. *Hum. Brain Mapp.* 33, 753–762. doi: 10.1002/hbm.21251

Gross, J. J. (2015). Emotion regulation: current status and future prospects. *Psychol. Inq.* 26, 1–26. doi: 10.1080/1047840X.2014.940781

Gundogdu, D., Finnerty, A. N., Staiano, J., Teso, S., Passerini, A., Pianesi, F., et al. (2017). Investigating the association between social interactions and personality states dynamics. *R. Soc. Open Sci.* 4:170194. doi: 10.1098/rsos.170194

Harbaugh, W. T., Mayr, U., and Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316, 1622–1625. doi: 10.1126/science.1140738

Henry, D., Dymnicki, A. B., Mohatt, N., Allen, J., and Kelly, J. G. (2015). Clustering methods with qualitative data: a mixed methods approach for prevention research with small samples. *Prev. Sci.* 16, 1007–1016. doi: 10.1007/s11121-015-0561-z

Heym, N., Firth, J., Kibowski, F., Sumich, A., Egan, V., and Bloxsom, C. A. J. (2019). Empathy at the heart of darkness: empathy deficits that bind the dark triad and those that mediate indirect relational aggression. *Front. Psych.* 10:95. doi: 10.3389/fpsyt.2019.00095

Hillis, A. E. (2014). Inability to empathize: brain lesions that disrupt sharing and understanding another's emotions. *Brain* 137, 981–997. doi: 10.1093/brain/awt317

Jones, D. N., and Neria, A. L. (2015). The dark triad and dispositional aggression. *Personal. Individ. Differ.* 86, 360–364. doi: 10.1016/j.paid.2015.06.021

Jones, D. N., and Paulhus, D. L. (2009). "Machiavellianism" in *Handbook of individual differences in social behavior* (New York: The Guilford Press), 93–108.

Jones, D. N., and Paulhus, D. L. (2011). The role of impulsivity in the dark triad of personality. *Personal. Individ. Differ.* 51, 679–682. doi: 10.1016/j.paid.2011.04.011

Kam, C. C. S., and Zhou, M. (2016). Is the Dark Triad Better Studied Using a Variable- or a Person-Centered Approach? An Exploratory Investigation. *PLoS One* 11:e0161628. doi: 10.1371/journal.pone.0161628

Kirk, U., Gu, X., Sharp, C., Hula, A., Fonagy, P., and Montague, P. R. (2016). Mindfulness training increases cooperative decision making in economic exchanges: evidence from fMRI. *Neuroimage* 138, 274–283. doi: 10.1016/j.neuroimage.2016.05.075

Kornilova, T. V., Kornilov, S., Chumakova, M., and Talmach, M. (2015). The dark triad personality traits measure: approbation of the dirty dozen questionnaire. *Psikholog. Zh.* 36, 99–112.

Korponay, C., Pujara, M., Deming, P., Philippi, C., Decety, J., Kosson, D. S., et al. (2017). Impulsive-antisocial dimension of psychopathy linked to enlargement and abnormal functional connectivity of the striatum. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 2, 149–157. doi: 10.1016/j.bpsc.2016.07.004

Lam, B. Y. H., Yang, Y., Schug, R. A., Han, C., Liu, J., and Lee, T. M. C. (2017). Psychopathy moderates the relationship between orbitofrontal and striatal alterations and violence: the investigation of individuals accused of homicide. *Front. Hum. Neurosci.* 11:579. doi: 10.3389/fnhum.2017.00579

Lämmle, L., and Ziegler, M. (2014). Dark triad – (anti) social behaviour to others and self. *Personal. Individ. Differ.* 60:S16. doi: 10.1016/j.paid.2013.07.371

Leopold, A., Krueger, F., dal Monte, O., Pardini, M., Pulaski, S. J., Solomon, J., et al. (2012). Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. *Soc. Cogn. Affect. Neurosci.* 7, 871–880. doi: 10.1093/scan/nsr071

Levy, K. N., Ellison, W. D., and Reynoso, J. S. (2011). A historical review of narcissism and narcissistic personality. In W. K. Campbell and J. D. Miller (Eds.), *The handbook of narcissism and narcissistic personality disorder: Theoretical approaches, empirical findings, and treatments* (pp. 3–13). Hoboken: John Wiley & Sons, Inc., 1–13.

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., and Tan, K. M. (2019). Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): causal, multivariate, and reverse inference evidence. *Neurosci. Biobehav. Rev.* 99, 311–328. doi: 10.1016/j.neubiorev.2018.12.021

Malesza, M., and Kalinowski, K. (2021). Dark triad and impulsivity – an ecological momentary assessment approach. *Curr. Psychol.* 40, 3682–3690. doi: 10.1007/s12144-019-00320-y

Maliske, L., and Kanske, P. (2022). The social connectome-moving toward complexity in the study of brain networks and their interactions in social cognitive and affective neuroscience. *Front. Psych.* 13:845492. doi: 10.3389/fpsyt.2022.845492

Maneiro, L., Navas, M. P., Van Geel, M., Cutrín, O., and Vedder, P. (2020). Dark triad traits and risky Behaviours: identifying risk profiles from a person-centred approach. *Int. J. Environ. Res. Public Health* 17:6194. doi: 10.3390/ijerph17176194

Massa, A. A., and Eckhardt, C. I. (2017). "Self-report psychopathy scale" in *Encyclopedia of personality and individual differences*. eds. V. Zeigler-Hill and T. K. Shackelford (Cham: Springer International Publishing), 1–4.

McLarnon, M. J. W. (2022). Into the heart of darkness: a person-centered exploration of the dark triad. *Personal. Individ. Differ.* 186:111354. doi: 10.1016/j.paid.2021.111354

Miao, C., Humphrey, R. H., Qian, S., and Pollack, J. M. (2019). The relationship between emotional intelligence and the dark triad personality traits: a meta-analytic review. *J. Res. Pers.* 78, 189–197. doi: 10.1016/j.jrp.2018.12.004

Michels, M., and Schulze, R. (2021). Emotional intelligence and the dark triad: a meta-analysis. *Personal. Individ. Differ.* 180:110961. doi: 10.1016/j.paid.2021.110961

Molenberghs, P., Johnson, H., Henry, J. D., and Mattingley, J. B. (2016). Understanding the minds of others: a neuroimaging meta-analysis. *Neurosci. Biobehav. Rev.* 65, 276–291. doi: 10.1016/j.neubiorev.2016.03.020

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., and Grafman, J. (2006). Human fronto–mesolimbic networks guide decisions about charitable donation. *Proc. Natl. Acad. Sci.* 103, 15623–15628. doi: 10.1073/pnas.0604475103

Morelli, S. A., Rameson, L. T., and Lieberman, M. D. (2014). The neural components of empathy: predicting daily prosocial behavior. *Soc. Cogn. Affect. Neurosci.* 9, 39–47. doi: 10.1093/scan/nss088

Moshagen, M., Hilbig, B. E., and Zettler, I. (2018). The dark core of personality. *Psychol. Rev.* 125, 656–688. doi: 10.1037/rev0000111

Muris, P., Merckelbach, H., Otgaar, H., and Meijer, E. (2017). The malevolent side of human nature: a meta-analysis and critical review of the literature on the dark triad (narcissism, Machiavellianism, and psychopathy). *Perspect. Psychol. Sci.* 12, 183–204. doi: 10.1177/1745691616666070

Myznikov, A., Zheltyakova, M., Korotkov, A., Kireev, M., Masharipov, R., Jagmurov, O. D., et al. (2021). Neuroanatomical correlates of social intelligence measured by the Guilford test. *Brain Topogr.* 34, 337–347. doi: 10.1007/s10548-021-00837-1

Nejati, V., Majdi, R., Salehinejad, M. A., and Nitsche, M. A. (2021). The role of dorsolateral and ventromedial prefrontal cortex in the processing of emotional dimensions. *Sci. Rep.* 11:1971. doi: 10.1038/s41598-021-81454-7

Nenadic, I., Güllmar, D., Dietzek, M., Langbein, K., Steinke, J., and Gaser, C. (2015). Brain structure in narcissistic personality disorder: a VBM and DTI pilot study. *Psychiatry Res. Neuroimaging* 231, 184–186. doi: 10.1016/j.pscychresns.2014.11.001

Nenadić, I., Lorenz, C., and Gaser, C. (2021). Narcissistic personality traits and prefrontal brain structure. *Sci. Rep.* 11:15707. doi: 10.1038/s41598-021-94920-z

Nickl-Jockschat, T., Habel, U., Maria Michel, T., Manning, J., Laird, A. R., Fox, P. T., et al. (2011). Brain structure anomalies in autism spectrum disorder—a meta-analysis of VBM studies using anatomic likelihood estimation. *Hum. Brain Mapp.* 33, 1470–1489. doi: 10.1002/hbm.21299

O'Connell, L. A., and Hofmann, H. A. (2011). The vertebrate mesolimbic reward system and social behavior network: a comparative synthesis. *J. Comp. Neurol.* 519, 3599–3639. doi: 10.1002/cne.22735

O'Sullivan, M., and Guilford, J. (1976). *Four factor tests of social intelligence (behavioral cognition): Manual of instructions and interpretations*. Orange, CA: Sheridan Psychological Services.

Pailing, A., Boon, J., and Egan, V. (2014). Personality, the dark triad and violence. *Personal. Individ. Differ.* 67, 81–86. doi: 10.1016/j.paid.2013.11.018

Pajevic, M., Vukosavljevic-Gvozden, T., Stevanovic, N., and Neumann, C. S. (2018). The relationship between the dark tetrad and a two-dimensional view of empathy. *Personal. Individ. Differ.* 123, 125–130. doi: 10.1016/j.paid.2017.11.009

Pan, N., Wang, S., Zhao, Y., Lai, H., Qin, K., Li, J., et al. (2021). Brain gray matter structures associated with trait impulsivity: a systematic review and voxel-based meta-analysis. *Hum. Brain Mapp.* 42, 2214–2235. doi: 10.1002/hbm.25361

Patrick, C. J. (2022). Psychopathy: current knowledge and future directions. *Annu. Rev. Clin. Psychol.* 18, 387–415. doi: 10.1146/annurev-clinpsy-072720-012851

Paulhus, D. L., and Williams, K. M. (2002). The dark triad of personality: narcissism, Machiavellianism, and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/S0092-6566(02)00505-6

Pechorro, P., Curtis, S., DeLisi, M., Maroco, J., and Nunes, C. (2022). Dark triad psychopathy outperforms self-control in predicting antisocial outcomes: a structural equation modeling approach. *Eur. J. Investig. Health Psychol. Educ.* 12, 549–562. doi: 10.3390/ejihpe12060041

Persson, B. N., Kajonius, P. J., and Garcia, D. (2019). Revisiting the structure of the short dark triad. *Assessment* 26, 3–16. doi: 10.1177/1073191117701192

Petrovic, P., Ekman, C. J., Klahr, J., Tigerström, L., Rydén, G., Johansson, A. G. M., et al. (2016). Significant grey matter changes in a region of the orbitofrontal cortex in healthy participants predicts emotional dysregulation. *Soc. Cogn. Affect. Neurosci.* 11, 1041–1049. doi: 10.1093/scan/nsv072

Riva, D., Bulgheroni, S., Aquino, D., Di Salle, F., Savoiardo, M., and Erbetta, A. (2011). Basal forebrain involvement in low-functioning autistic children: a voxel-based morphometry study. *AJNR Am. J. Neuroradiol.* 32, 1430–1435. doi: 10.3174/ajnr.A2527

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Schulz, J., Brandl, F., Grothe, M. J., Kirschner, M., Kaiser, S., Schmidt, A., et al. (2023). Basal-forebrain cholinergic nuclei alterations are associated with medication and cognitive deficits across the schizophrenia Spectrum. *Schizophr. Bull.* 49, 1530–1541. doi: 10.1093/schbul/sbad118

Segel-Karpas, D., and Lachman, M. E. (2018). Social contact and cognitive functioning: the role of personality. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 73, gbw079–gbw984. doi: 10.1093/geronb/gbw079

Shamay-Tsoory, S. G., and Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia* 45, 3054–3067. doi: 10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., and Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Soc. Neurosci.* 1, 149–166. doi: 10.1080/17470910600985589

Sijtsema, J. J., Garofalo, C., Jansen, K., and Klimstra, T. A. (2019). Disengaging from evil: longitudinal associations between the dark triad, moral disengagement, and antisocial behavior in adolescence. *J. Abnorm. Child Psychol.* 47, 1351–1365. doi: 10.1007/s10802-019-00519-4

Trahair, C., Baran, L., Flakus, M., Kowalski, C. M., and Rogoza, R. (2020). The structure of the dark triad traits: a network analysis. *Personal. Individ. Differ.* 167:110265. doi: 10.1016/j.paid.2020.110265

Truhan, T. E., Wilson, P., Mõttus, R., and Papageorgiou, K. A. (2021). The many faces of dark personalities: an examination of the dark triad structure using psychometric network analysis. *Personal. Individ. Differ.* 171:110502. doi: 10.1016/j.paid.2020.110502

Verbeke, W. J. M. I., Rietdijk, W. J. R., van den Berg, W. E., Dietvorst, R. C., Worm, L., and Bagozzi, R. P. (2011). The making of the Machiavellian brain: a structural MRI analysis. *J. Neurosci. Psychol. Econ.* 4, 205–216. doi: 10.1037/a0025802

Vieira, J. B., Ferreira-Santos, F., Almeida, P. R., Barbosa, F., Marques-Teixeira, J., and Marsh, A. A. (2015). Psychopathic traits are associated with cortical and subcortical volume alterations in healthy individuals. *Soc. Cogn. Affect. Neurosci.* 10, 1693–1704. doi: 10.1093/scan/nsv062

Votinov, M., Myznikov, A., Zheltyakova, M., Masharipov, R., Korotkov, A., Cherednichenko, D., et al. (2021). The interaction between caudate nucleus and regions within the theory of mind network as a neural basis for social intelligence. *Front. Neural Circuits* 15:727960. doi: 10.3389/fncir.2021.727960

Walker, S. A., Olderbak, S., Gorodezki, J., Zhang, M., Ho, C., and Mac Cann, C. (2022). Primary and secondary psychopathy relate to lower cognitive reappraisal: a meta-analysis of the dark triad and emotion regulation processes. *Personal. Individ. Differ.* 187:111394. doi: 10.1016/j.paid.2021.111394

Yang, Y., and Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Res.* 174, 81–88. doi: 10.1016/j.pscychresns.2009.03.012

Yang, Y., Raine, A., Colletti, P., Toga, A. W., and Narr, K. L. (2010). Morphological alterations in the prefrontal cortex and the amygdala in unsuccessful psychopaths. *J. Abnorm. Psychol.* 119, 546–554. doi: 10.1037/a0019611

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 14 March 2024
DOI 10.3389/fpsyg.2024.1271916

OPEN ACCESS

EDITED BY
Chiara Lucifora,
University of Bologna, Italy

REVIEWED BY
Valerio Capraro,
Middlesex University, United Kingdom
Hansika Singhal,
University of Delhi, India

*CORRESPONDENCE
Yu-Wei Wu
✉ 986958226@qq.com
Zi-Ye Huang
✉ 18106790817@163.com

RECEIVED 03 August 2023
ACCEPTED 26 February 2024
PUBLISHED 14 March 2024

CITATION
Li H-M, Yan W-J, Wu Y-W and Huang Z-Y (2024) Cognitive control in honesty and dishonesty under different conflict scenarios: insights from reaction time.
Front. Psychol. 15:1271916.
doi: 10.3389/fpsyg.2024.1271916

Cognitive control in honesty and dishonesty under different conflict scenarios: insights from reaction time

Hao-Ming Li[1], Wen-Jing Yan[2], Yu-Wei Wu[3]* and Zi-Ye Huang[1]*

[1]Wenzhou Seventh People's Hospital, Wenzhou, China, [2]School of Mental Health, Wenzhou Medical University, Wenzhou, China, [3]Student Affairs Division, Wenzhou Business College, Wenzhou, China

This study investigated the role of cognitive control in moral decision-making, focusing on conflicts between financial temptations and the integrity of honesty. We employed a perceptual task by asking participants to identify which side of the diagonal contained more red dots within a square to provoke both honest and dishonest behaviors, tracking their reaction times (RTs). Participants encountered situations with no conflict, ambiguous conflict, and clear conflict. Their behaviors in the clear conflict condition categorized them as either "honest" or "dishonest." Our findings suggested that, in ambiguous conflict situations, honest individuals had significantly longer RTs and fewer self-interest responses than their dishonest counterparts, suggesting a greater need for cognitive control to resolve conflicts and a lesser tendency toward self-interest. Moreover, a negative correlation was found between participants' number of self-interest responses and RTs in ambiguous conflict situations ($r = -0.27$ in study 1 and $r = -0.66$ in study 2), and a positive correlation with cheating numbers in clear conflict situations ($r = 0.36$ in study 1 and $r = 0.82$ in study 2). This suggests less cognitive control was required for self-interest and cheating responses, bolstering the "Will" hypothesis. We also found that a person's self-interest tendency could predict their dishonest behavior. These insights extend our understanding of the role of cognitive control plays in honesty and dishonesty, with potential applications in education, policy-making, and business ethics.

KEYWORDS

honesty, dishonesty, cognitive control, moral decision-making, reaction time

1 Introduction

Human behavior is often governed by complex decision-making processes, with one recurring challenge being the conflict between self-interest and the pursuit of moral righteousness. This moral quandary, the struggle between the temptation of personal financial gain and the aspiration to uphold an honest image, unfolds in various scenarios ranging from relatively minor instances of tax evasion and inflated expense reports, to more severe instances of fraudulent financial schemes (Mazar et al., 2008). Such moral dilemmas offer a fascinating window into human behavior and motivations. They invite questions regarding how individuals reconcile these seemingly incompatible drives of personal gain and moral obligation. An increasingly explored proposition within the behavioral sciences is that cognitive control, our inherent ability to regulate thoughts,

emotions, and actions, acts as a mediator in this tension between self-interest and moral self-image (Wood et al., 2008).

Despite its intuitive appeal, the role of cognitive control in moral decision-making, particularly its contribution to resolving conflicts between self-interest and honesty, remains a contentious topic within psychological and neuroscientific research. Although a large amount of data is available, the results are mixed (Köbis et al., 2019; Capraro, 2023). This debate predominantly centers around two main hypotheses: the "Will" hypothesis and the "Grace" hypothesis (Tabatabaeian et al., 2015). The "Will" hypothesis paints a less flattering image of human nature. It posits that humans, by default, are selfish and dishonest, and that it is cognitive control that keeps these basic instincts in check, compelling individuals toward honesty (Gino et al., 2011). This hypothesis aligns with traditional economic models of human behavior, which suggest that individuals are naturally driven to pursue self-interest, with social norms, laws, and moral values acting as external constraints on these inborn desires (Becker, 1968; Henrich et al., 2005). Contrastingly, the "Grace" hypothesis presents a more favorable image of humans, suggesting that people are essentially honest, and that cognitive control is used to suppress instinctual honest responses when there are opportunities to profit from dishonesty (Rand et al., 2012). This view is supported by empirical research that shows individuals respond faster when instructed to tell the truth than when directed to lie, suggesting honesty may indeed be more intuitive (Capraro, 2017; Suchotzki et al., 2017; Verschuere et al., 2018; Capraro et al., 2019).

This ongoing debate is far from a mere academic exercise. Instead, it underscores the complex, multifaceted nature of human morality, highlighting the need for more nuanced and empirical investigations into the interplay between cognitive control and moral behavior (Baumeister et al., 2009). As Baumeister and Exline (1999) propose, understanding these moral dynamics requires acknowledging individual differences, considering situational variables, and appreciating the dynamic nature of moral decision-making processes. Study found that the social consequences of lying could be a promising key to the riddle of intuition's role in honesty. When dishonesty harms abstract others, promoting intuition causes more people to lie and people to lie more. However, when dishonesty harms concrete others, promoting intuition has no significant effect on dishonesty (Köbis et al., 2019). Recent research advancements have further complicated this landscape. With the advent of neuroimaging techniques, studies suggest that the impact of cognitive control on moral behavior may be dependent on an individual's inherent moral disposition toward honesty or dishonesty (Greene and Paxton, 2009). Specifically, brain regions associated with cognitive control, such as the anterior cingulate cortex and the inferior frontal gyrus, have been found to help individuals predisposed toward dishonesty to act honestly, while enabling those predisposed toward honesty to cheat when the situation permit (Speer et al., 2022).

Against this backdrop, the present study embarks on an exploration of how individuals, predisposed toward honesty or dishonesty, respond to situations that present a conflict between personal financial gain and moral self-image. Beyond neuroimaging, reaction time (RT) measures, often utilized in cognitive psychology, are believed to offer critical insights into cognitive control's involvement in moral decision-making conflicts (Evans et al., 2015; Andrighetto et al., 2020). RTs provide non-invasive, real-time evidence of the cognitive processes at play during moral decision-making (Shalvi et al., 2012). This study introduces three distinct decision conflict scenarios, allowing for a more nuanced examination of individual differences in cognitive control and moral tendencies. By analyzing the interaction between cognitive control, moral inclination, and response times across these scenarios, we hope to provide a more comprehensive, more dynamic, and ultimately, a more human perspective on the landscape of moral decision-making (Tangney et al., 2007).

2 Study 1

2.1 Method

2.1.1 Participants

We recruited sixty-seven undergraduate or postgraduate students from Wenzhou University. The participants, with an average age of 19.39 ($SD = 1.18$), comprised 39 females and 28 males. All participants had normal or corrected-to-normal vision and provided informed consent before the study. The study adhered to the sixth revision of the Declaration of Helsinki (2008) and was approved by the university's Institutional Review Board (IRB).

2.1.2 Procedure

We engaged the participants in a perceptual task (Gino et al., 2010). Each trial presented a square divided diagonally into two sections. Each section had 20 dots scattered randomly on the left or right side of the diagonal. After a one-second exposure, participants identified which side of the diagonal held more dots by clicking the respective mouse button. The reward for each trial was calculated as follows: clicking the left mouse button yielded 0.02CNY, whereas clicking the right button yielded 0.2CNY. Therefore, trials with more dots on the left side of the diagonal presented a conflict between answering accurately and maximizing profit.

The perceptual task was split into two phases. The first phase consisted of 100 practice trials, after which participants received feedback on potential earnings for each trial and cumulative earnings if these trials had involved real payment. In the second phase, the participants completed 200 trials, earning real money, and received information about their earnings for each trial and overall.

Participants could earn a maximum of 40CNY on this perceptual task (by always pressing the right mouse button). There were four blocks. Each block consisted of 50 trials, and each block included 8 trials in which the answer was clearly "more on right" (no conflict condition, i.e., the ratio of the number of dots on the right to the number of dots on the left was greater than or equal to 1.5), 17 trials in which the answer was clearly "more on left" (clear conflict condition, i.e., the ratio of the number of dots on the right to the number of dots on the left was less than or equal to 2/3), and 25 ambiguous trials (ambiguous conflict condition, i.e., the ratio of the number of dots on the right to the number of dots on the left was between 2/3 and 1.5). The responses in ambiguous condition reflect an individual's self-interest tendency. Once participants completed this task, the computer indicated that they should report their performance in Phase 2 on a collection slip to be handed to the experimenter at the end of the study.

2.2 Results

All participants displayed honest behavior in no conflict condition. A total of 53.73% (36/67) participants were found to cheat one or more times in clear conflict condition (see Figure 1). The participants who cheated in clear conflict condition (36 participants, mean cheating number = 23.63) will be referred to as 'dishonest individuals', while the remaining participants (31 participants) will be referred to as 'honest individuals'.

We compared the reaction times (RTs) in the no conflict, ambiguous conflict and clear conflict conditions among honest and dishonest participants. Results showed that honest individuals required longer RTs than dishonest individuals in the ambiguous conflict condition, $p = 0.047$, suggesting that honest individuals required more time to revolve ambiguous conflict. Also, honest individuals made less self-interest responses ($M = 42$, $SD = 5.08$) than dishonest individuals ($M = 67.42$, $SD = 22.34$) in the ambiguous condition, $p < 0.001$. There were no RT differences in no conflict and clear conflict conditions among honest and dishonest participants, $p = 0.07$; $p = 0.09$.

Moreover, the RTs in ambiguous trials correlated with the self-interest numbers in the ambiguous condition, $r = -0.27$, $p = 0.028$ and the cheating numbers in the clear conflict condition, $r = -0.24$, $p = 0.046$. The self-interest numbers in the ambiguous condition correlated with the cheating numbers in the clear conflict condition, $r = 0.36$, $p = 0.003$. When using RTs and self-interest numbers in the ambiguous condition to predict the cheating numbers in the clear conflict condition, the model was significant, with R2 = 0.15, $p = 0.005$. The self-interest number in the ambiguous condition was a significant indicator of cheating numbers in the clear conflict condition, $p = 0.01$; whereas the RTs in the ambiguous conditions was not significant in the model, $p = 0.19$.

3 Study 2

3.1 Method

3.1.1 Participant

We recruited ninety-five undergraduate or postgraduate students from Hebei Normal University. The participants, with an average age of 19.55 ($SD = 1.07$), comprised 75 females and 20 males. All participants had normal or corrected-to-normal vision and provided informed consent before the study. The study adhered to the sixth revision of the Declaration of Helsinki (2008) and was approved by the university's Institutional Review Board (IRB).

3.1.2 Procedure

The task is same as that of Study 1, only some differences in experimental materials. In the experiment, 18 images were made in the order of the left and right red dots from less to most. The experiment consisted of 4 blocks, and 18 images in each block were randomly presented 4 times, for a total of 72 trials. The experiment consisted of a total of 288 trials.

3.2 Results

All participants displayed honest behavior in no conflict condition. A total of 54.74% (52/95) participants were found to cheat one or more times in clear conflict condition. The participants who cheated in clear conflict condition (52 participants, mean cheating number = 29.98) will be referred to as 'dishonest individuals', while the remaining participants (43 participants) will be referred to as 'honest individuals'.

We compared the reaction times (RTs) in the no conflict, ambiguous conflict and clear conflict conditions among honest and dishonest participants. Results showed that honest individuals required longer RTs ($M = 665.25$, $SD = 143.04$) than dishonest individuals in ($M = 550.64$, $SD = 166.12$) the ambiguous conflict condition, $p = 0.001$, suggesting that honest individuals required more time to revolve ambiguous conflict. Also, honest individuals made less self-interest responses ($M = 3.50$, $SD = 2.48$) than dishonest individuals ($M = 9.69$, $SD = 4.78$) in the ambiguous condition, $p < 0.001$. Moreover, the RTs in ambiguous trials correlated with the self-interest numbers in the ambiguous condition, $r = -0.66$, $p < 0.001$ and the cheating numbers in the clear conflict condition, $r = -0.65$, $p < 0.001$. The self-interest numbers in the ambiguous condition correlated with the cheating numbers in the clear conflict condition, $r = 0.82$, $p < 0.001$. When using RTs and self-interest numbers in the ambiguous condition to predict the cheating numbers in the clear conflict condition, the

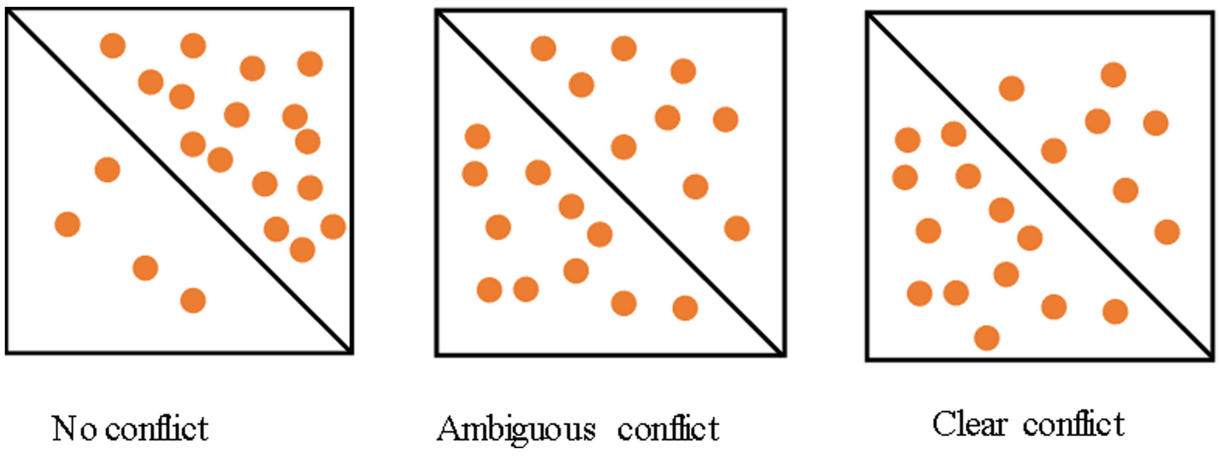


FIGURE 1
Examples of decision conflicts.

model was significant, with $R^2 = 0.73$, $p < 0.001$. The self-interest number and the RTs in the ambiguous conditions were significant indicators of cheating numbers in the clear conflict condition, $p < 0.001$; $p = 0.003$.

We also investigated the effect of conflict degree on the RTs of honest and dishonest people. Subtract the non-conflicting RTs from the conflicting RTs corresponding to the left and right red dots (i.e., the RTs under the condition that the left red dot is 13 minus the RTs under the condition that the right red dot is 7; The RTs under the condition that the left red dot is 14 minus the RTs under the condition that the red dot on the right is 6; and so on). We believe that the smaller difference between the numbers of red dots on the left and right, the greater psychological conflict of the individual. The results showed that conflict degree affected the participants' responses, the greater the conflict, the longer RTs required, $F(6, 498) = 47.67$, $p < 0.001$. There was no difference between the honest and dishonest people in their RTs at different conflict levels, $p = 0.46$.

4 Discussion

Our study investigated the interplay between cognitive control and moral decision-making, particularly focusing on how individuals with different predispositions toward honesty or dishonesty react in situations where personal financial gain conflicts with moral self-image. The key finding is that individuals who are inherently more honest exhibited longer reaction times in scenarios with ambiguous moral conflicts, suggesting a deeper cognitive engagement in these dilemmas. Conversely, those predisposed to dishonesty responded more quickly, implying less cognitive deliberation. This differentiation highlights the complex role of cognitive control in navigating moral decisions, indicating that it is influenced by an individual's moral inclinations. Essentially, our results contribute to understanding the nuanced mechanisms behind moral behavior, showing that moral decision-making is a dynamic process shaped by both cognitive control and personal ethical standards.

Our observation that honest individuals exhibit longer reaction times in ambiguous conflict conditions than their dishonest counterparts offers an intriguing insight into the cognitive processes underlying moral behavior. This finding aligns with the work of Capraro and Rand (2018), who suggested that honesty might be more intuitive to individuals with a stronger predisposition toward prosocial behavior, requiring less cognitive control in clear-cut situations but more deliberation when the context is ambiguous. Our results extend this theory by quantitatively showing that the cognitive effort, as measured by reaction times, increases in moral dilemmas where the right choice is not immediately apparent.

Additionally, the correlation between reaction times and self-interest behaviors in ambiguous and clear conflict conditions, as observed in our study, indicates a dynamic interplay between cognitive control and situational factors. This extends the findings of Shalvi et al. (2012), who highlighted the role of situational clarity in ethical decision-making. Our results further elaborate on this by showing that the ambiguity of a situation not only affects decision-making speed but also interacts with an individual's moral inclination to influence their choices.

Furthermore, our study contributes to the debate surrounding the "Will" and "Grace" hypotheses. The negative correlation between cognitive control and the number of self-interest responses suggests that honesty, far from being the default human condition, may be the product of a conscious cognitive effort to restrain self-serving impulses. This would be consistent with the "Will" hypothesis.

4.1 Applications and limitations

The results extend our understanding of the role of cognitive control plays in honesty and dishonesty, with potential applications in education, policy-making, and business ethics. For educational settings, the results suggest curricula should emphasize enhancing ethical reasoning and cognitive control, preparing students to navigate moral challenges thoughtfully. Policy implications include designing environments that discourage dishonesty by clarifying ethical standards and making dishonest actions more cognitively taxing, thereby promoting transparency and accountability. In business ethics, our findings advocate for cultures of integrity supported by clear ethical guidelines and training programs that bolster moral awareness and cognitive control, helping employees prioritize ethical standards over self-interest. This approach aims to foster a more honest and ethical conduct across various sectors.

Our study, while offering valuable insights into the complex interplay between cognitive control and moral decision-making, is not without its limitations. One of the primary constraints involves the sample size and demographic composition, primarily undergraduate and postgraduate students, which may not fully represent the broader population. This limitation could affect the generalizability of our findings, as the specific age group and educational background of our participants might influence their moral decision-making processes and cognitive control mechanisms differently compared to a more diverse population. Additionally, our reliance on reaction times as the use of intuitive or reflective processes should be careful. Rather some studies highlight the pitfalls of using RT correlations as support for dual-process theories. Reaction times, in this context, primarily reflect the cognitive processing involved in navigating moral conflicts rather than directly indicating whether honesty is an inherent or automatic response (Evans et al., 2015; Krajbich et al., 2015; Andrighetto et al., 2020).

5 Conclusion

Our study contributes to the nuanced understanding of the interplay between cognitive control and moral decision-making, revealing the complex mechanisms through which individuals navigate ethical dilemmas. By examining the roles of decision conflict and moral deliberation across different moral predispositions, our findings challenge and extend existing theories on moral psychology. Despite limitations related to sample diversity and the interpretation of reaction times, this research underscores the importance of considering individual differences and the multifaceted nature of cognitive processes in ethical behavior. Looking forward, it paves the way for further interdisciplinary investigations into moral decision-making, encouraging a broader exploration of how cognitive, emotional, and social factors collectively shape our moral actions. As we continue to unravel the cognitive underpinnings of morality, this work not only deepens our theoretical understanding but also has practical implications for promoting ethical behavior in an increasingly complex world.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by IRB of the Seventh People's Hospital of Wenzhou (EC-KY-2022048). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

H-ML: Writing – original draft. W-JY: Conceptualization, Writing – original draft. Y-WW: Conceptualization, Writing – original draft. Z-YH: Writing – review & editing, Conceptualization, Investigation, Validation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by Wenzhou Science and Technology Project (Y2023864).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andrighetto, G., Capraro, V., Guido, A., and Szekely, A. (2020). Cooperation, response time, and social value orientation: a meta-analysis. *Proc. Cogn. Sci. Soc.*, 2116–2122. doi: 10.31234/osf.io/cbakz

Baumeister, R. F., and Juola Exline, J. (1999). Virtue, personality, and social relations: Self-control as the moral muscle. *Journal of personality*, 67, 1165–1194. doi: 10.1111/1467-6494.00086

Baumeister, R. F., Masicampo, E. J., and DeWall, C. N. (2009). Prosocial benefits of feeling free: disbelief in free will increases aggression and reduces helpfulness. *Personal. Soc. Psychol. Bull.* 35, 260–268. doi: 10.1177/0146167208327217

Becker, G. S. (1968). Crime and punishment: an economic approach. *J. Polit. Econ.* 76, 169–217. doi: 10.1086/259394

Capraro, V. (2017). Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Econ. Lett.* 158, 54–57. doi: 10.1016/j.econlet.2017.06.015

Capraro, V. (2023). The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *arXiv*. doi: 10.48550/arXiv.1906.09948

Capraro, V., and Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making*, 13, 99–111. doi: 10.1017/S1930297500008858

Capraro, V., Schulz, J., and Rand, D. G. (2019). Time pressure and honesty in a deception game. *J. Behav. Exp. Econ.* 79, 93–99. doi: 10.1016/j.socec.2019.01.007

Evans, A. M., Dillon, K. D., and Rand, D. G. (2015). Fast but not intuitive, slow but not reflective: decision conflict drives reaction times in social dilemmas. *J. Exp. Psychol. Gen.* 144, 951–966. doi: 10.1037/xge0000107

Gino, F., Norton, M. I., and Ariely, D. (2010). The counterfeit self: the deceptive costs of faking it. *Psychol. Sci.* 21, 712–720. doi: 10.1177/0956797610366545

Gino, F., Schweitzer, M. E., Mead, N. L., and Ariely, D. (2011). Unable to resist temptation: how self-control depletion promotes unethical behavior. *Organ. Behav. Hum. Decis. Process.* 115, 191–203. doi: 10.1016/j.obhdp.2011.03.001

Greene, J. D., and Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci.* 106, 12506–12511. doi: 10.1073/pnas.0900152106

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behav. Brain Sci.* 28, 795–815. doi: 10.1017/S0140525X05000142

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspect. Psychol. Sci.* 14, 778–796. doi: 10.1177/1745691619851778

Krajbich, I., Bartling, B., Hare, T., and Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat. Commun.* 6:7455. doi: 10.1038/ncomms8455

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* 45, 633–644. doi: 10.1509/jmkr.45.6.633

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature* 489, 427–430. doi: 10.1038/nature11467

Shalvi, S., Eldar, O., and Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychol. Sci.* 23, 1264–1270. doi: 10.1177/0956797612443835

Speer, S. P., Smidts, A., and Boksem, M. A. (2022). Cognitive control and dishonesty. *Trends Cogn. Sci.* 26, 796–808. doi: 10.1016/j.tics.2022.06.005

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., and Crombez, G. (2017). Lying takes time: a Meta-analysis on reaction time measures of deception. *Psychol. Bull.* 143, 428–453. doi: 10.1037/bul0000087

Tabatabaeian, M., Dale, R., and Duran, N. D. (2015). Self-serving dishonest decisions can show facilitated cognitive dynamics. *Cogn. Process.* 16, 291–300. doi: 10.1007/s10339-015-0660-6

Tangney, J. P., Stuewig, J., and Mashek, D. J. (2007). Moral emotions and moral behavior. *Annu. Rev. Psychol.* 58, 345–372. doi: 10.1146/annurev.psych.56.091103.070145

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2018). Taxing the brain to uncover lying? Meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying. *J. Appl. Res. Mem. Cogn.* 7, 462–469. doi: 10.1016/j.jarmac.2018.04.005

Wood, A. M., Linley, P. A., Maltby, J., Baliousis, M., and Joseph, S. (2008). The authentic personality: a theoretical and empirical conceptualization and the development of the authenticity scale. *J. Couns. Psychol.* 55, 385–399. doi: 10.1037/0022-0167.55.3.385

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 19 March 2024
DOI 10.3389/fpsyg.2024.1335548

OPEN ACCESS

EDITED BY
Carmelo Mario Vicario,
University of Messina, Italy

REVIEWED BY
Thomas Quettier,
University of Padua, Italy
Wenhai Zhang,
Hengyang Normal University, China

*CORRESPONDENCE
Qing Liu
✉ psylq@zjut.edu.cn

RECEIVED 09 November 2023
ACCEPTED 27 February 2024
PUBLISHED 19 March 2024

CITATION
Dai S, Liu Q, Chai H and Zhang W (2024) Neural mechanisms of different types of envy: a meta-analysis of activation likelihood estimation methods for brain imaging. *Front. Psychol.* 15:1335548. doi: 10.3389/fpsyg.2024.1335548

Neural mechanisms of different types of envy: a meta-analysis of activation likelihood estimation methods for brain imaging

Shuchang Dai[1], Qing Liu[1]*, Hao Chai[1] and Wenjuan Zhang[2]

[1]College of Education and Technology, Zhejiang University of Technology, Hangzhou, China, [2]Mental Health Education Center, Xidian University, Xi'an, Shaanxi, China

Previous studies have a lack of meta-analytic studies comparing the trait (personality) envy, social comparison envy, and love–envy, and the understanding of the similarities and differences in the neural mechanisms behind them is relatively unclear. A meta-analysis of activation likelihood estimates was conducted using 13 functional magnetic resonance imaging studies. Studies first used single meta-analyses to identify brain activation areas for the three envy types. Further, joint and comparative analyses were followed to assess the common and unique neural activities among the three envy types. A single meta-analysis showed that the critical brain regions activated by trait (personality) envy included the inferior frontal gyrus, cingulate gyrus, middle frontal gyrus, lentiform nucleus and so on. The critical brain regions activated by social comparison envy included the middle frontal gyrus, inferior frontal gyrus, medial frontal gyrus, precuneus and so on. The critical brain regions activated by love–envy included the inferior frontal gyrus, superior frontal gyrus, cingulate gyrus, insula and so on. In terms of the mechanisms that generate the three types of envy, each of them is unique when it comes to the perception of stimuli in a context; in terms of the emotion regulation mechanisms of envy, the three types of envy share very similar neural mechanisms. Both their generation and regulation mechanisms are largely consistent with the cognitive control model of emotion regulation. The results of the joint analysis showed that the brain areas co-activated by trait (personality) envy and social comparison envy were frontal sub-Gyral, inferior parietal lobule, inferior frontal gyrus, precuneus and so on; the brain areas co-activated by trait (personality) envy and love–envy were extra-nuclear lobule, lentiform nucleus, paracentral lobule, cingulate gyrus and so on; the brain regions that are co-activated by social comparison envy and love–envy are anterior cingulate gyrus, insula, supramarginal gyrus, inferior frontal gyrus and so on. The results of the comparative analysis showed no activation clusters in the comparisons of the three types of envy.

KEYWORDS

envy, functional magnetic resonance imaging, activation likelihood estimation, meta analyses, neural mechanism

1 Introduction

Envy is a psychological and behavioral activity prevalent in human societies. For individuals, it is a rich emotional experience. Envy is considered as a combination of the primary emotions anger, fear and sadness (Zheng et al., 2019). Along with the experience of an unpleasant emotional state, envy also is associated with a host of behaviors. Extreme,

pathological envy includes delusional symptoms and promotes aggression in terms of domestic violence, self-mutilation and even murder (Camicioli, 2011) and can also occur in association with depression and autism (Bauminger, 2004). Conversely, positive outcomes related to envy also have been reported, including motivating people to do better than their competitor (Protasi, 2016), by for example, inspiring individuals to improve their position in the workplace. For groups, envy is also a complex social culture and phenomenon (Pang, 2016). Envy usually occurs in social interactive contexts, such as sexual infidelity or social comparison scenarios. From an evolutionary perspective, envy prevents an individual from being outperformed by a direct competitor in a fitness-relevant domain: Envy motivates behaviors towards gaining a similar standing as a competitor or acting to remove a competitor's advantage. Therefore, We experience envy when the positive attributes of another individual jeopardize our social standing (Crusius and Lange, 2016).

The phenomenon of envy is complex, and to understand it more clearly, based on scientific research, psychologists have categorized envy. According to Bringle's (1991) categorization, envy can be broadly classified into two types - suspicious envy and reactive envy. Bringle's Interaction Model Theory of Envy states that envy reactions are the outcome of an interaction between endogenous (internal) and exogenous (external) variables, such as the environment and culture. However, the impact of these variables may differ in each individual, leading to different types of envy. When the internal variable plays a significant role in determining the envy response, it is known as suspicious envy. On the other hand, if the external variable is an important determinant, it is called reactive envy. Currently, the three types of envy commonly accepted by the general public and the subject of much research are the trait (personality) envy, social comparison envy, and love–envy. Of the three types of envy, suspicious envy is typified by trait (personality) envy. Outcomes from a relationship, comparison level (CL), and comparison level of alternatives (CLalt) are typically viewed as situational determinants of envy (Bringle, 1981). The Bringle Self-report Jealousy Scale (BSJS), which was developed by Bringle et al. (1979), is a tool that measures an individual's experience of self-envy in different contexts. It does so by using two dimensions - Social Comparison Envy and Love–Envy. It can be seen that reactive envy is typified by social comparison envy and love–envy. They all have their own unique characteristics.

The response of trait (personality) envy is mainly determined by endogenous variables, which are related to individuals. Envy is thought to arise from the perceived threat of losing respect and social status in the eyes of others (Silver and Sabini, 1978; Fiske, 2010; Crusius and Lange, 2016). Unlike situational envy, which manifests based on specific tasks, trait (personality) envy exhibits a general sensitivity to status threats. However, the tendency to react negatively emotionally and the corresponding behavioral changes vary across individuals. Empirical research has shown that individuals differ in the extent to which they desire social status (Anderson et al., 2015) and compare themselves to others (Gibbons and Buunk, 1999), and thus differ in their tendency to experience envy when faced with upward social comparison. Moreover, comparison-related personality disposition traits (Gibbons and Buunk, 1999) may shape trait (personality) envy, such as inequity aversion, justice sensitivity, and achievement motivation (Steinbeis and Singer, 2013; Lange and Crusius, 2015a). As for the measurement of the trait (personality) envy, researchers have independently developed and validated various scales, including the Envy subscale of the Materialism Scale developed by Belk (1985), the Multidimensional Jealousy Scale (MJS) developed by Pfeiffer and Wong (1989), the York Enviousness Scale (YES) developed by Gold (1996) from York University, Canada, the Dispositional Envy Scale (DES) developed by Smith et al. (1999), a nine-point envy scale developed by Lange and Crusius (2015b).

The response of social comparison envy and love–envy is mainly determined by exogenous variables, which are related to social and cultural. According to the theory of social comparison envy, social comparison is an essential aspect of social interaction. This process involves individuals comparing their beliefs, attitudes, and opinions with those of others (Festinger, 1954). However, when individuals engage in unfavorable upward social comparisons, they may experience painful feelings of envy (Silver and Sabini, 1978; Salovey and Rodin, 1984). Furthermore, behavioral research on envy has also confirmed that the more a person compares themselves to others, the more jealous experiences they will experience (Smith et al., 1999; Zeelenberg and Pieters, 2007). Love–envy is considered to be an emotion experienced when an individual faces the loss of an existing significant relationship with another person (meaning a companion) because of a third person (Mathes, 1992), which includes love–envy resulting from infidelity (sexual or emotional infidelity). Love–envy can sometimes have adverse effects, particularly triggering behaviors such as excessive snooping, controlling companions, and verbal or physical aggression (Kar and O'Leary, 2013; Neal and Lemay, 2014). These behaviors can damage intimate relationships between companions and may even lead to malignant events, such as domestic violence (Dandurand and Lafontaine, 2014; Deans and Bhogal, 2019). Social comparison envy and love–envy were the two most common types of reactive envy. The Bringle Self-Report Envy Scale (BSJS), developed by Bringle et al. (1979), includes both social comparison envy and love–envy dimensions to measure individuals' extensive experiences of self-envy in various contexts. In contrast, the Interpersonal Relationship Scale (IRS) developed by Hupka and Bachelor (1979) measured a single envy type.

Many persons view envy as neutral: It is neither only good nor only bad. Thus, elimination of all envy is not necessarily a desirable outcome. One should be prepared to cope with real, impending threats. Managing envy so that it becomes a constructive factor in a relationship is desirable. Thus, one should explore positive ways to cope with feelings. Self-management of envy is closely linked to emotion regulation. Emotion regulation includes a wide range of cognitive, behavioral, emotional, and physiological responses and is necessary to understand the emotional and behavioral correlates of stress and negative emotional states (Garnefski and Kraaij, 2006). Ochsner and Gross (2007) constructed a cognitive control model of emotion regulation with a bottom-up and top-down perspective. According to the theory, generating emotions involves four stages. In the first stage, a stimulus is perceived in its current situational context. At the second stage, one attends to some of these stimuli or their attributes. The third stage involves appraising the significance of stimuli in terms of their relevance to one's current goals, wants or needs. Finally, the fourth stage involves translating these appraisals into changes in experience, emotion-expressive behavior, and autonomic physiology (Ochsner and Gross, 2008). Cognitive reappraisal and expressive inhibition are two strategies for regulating emotions. The cognitive control model of emotion regulation suggests that emotion regulation arises during the process of emotion onset

and that different emotion regulation occurs at different stages of emotion onset (Gross and James, 1998; Gross and Thompson, 2007). Among them, cognitive changes are formed before the formation of emotional response tendencies, which are prior-focused emotion regulation and exhibit cognitive reappraisal emotion regulation strategies; response adjustments are made after the formation of emotional response tendencies, which are response-focused emotion regulation and exhibit expression inhibition of emotion regulation strategies.

Previous studies have addressed the neural mechanisms underlying envy less frequently, and only a very few studies have examined the neural mechanisms associated with non-pathological envy in healthy individuals. These researchers have used brain imaging techniques such as functional magnetic resonance imaging (fMRI) and other brain imaging techniques to explore the neural mechanisms underlying different types of envy, trying to find structural and functional markers associated with envy in the brain. In an fMRI study on trait (personality) envy, Xiang et al. (2016) used regional homogeneity (ReHo) to measure trait (personality) envy. They found that the inferior frontal gyrus (IFG), middle frontal gyrus (MFG), and dorsomedial prefrontal cortex (DMPFC) were found to be positive predictors of personality envy; Xiang et al. (2017) used a voxel-based morphometry (VBM) approach to measure trait (personality) envy and found that trait (personality) envy was positively correlated with dorsolateral prefrontal cortex (DLPFC) and superior temporal gyrus (STG) were positively correlated; Zheng et al. (2019) used a neural representation of emotions to measure trait (personality) envy and found that insula, fusiform gyrus (FG), hippocampus, dorsal striatum (DS), and inferior frontal gyrus (IFG) were found to have increased activation in these brain regions. In an fMRI study on social comparison envy, Dvash et al. (2010) found activation of the ventral striatum (*VS*) by inducing social comparison envy through a money gain or loss game; Tanaka et al. (2019) used a slightly different money gain or loss game than the former paradigm to induce social comparison envy and found that the dorsal anterior cingulate cortex (dACC) was activated; Brennan et al. (2020) used a story context approach to induce social comparison envy and found that the superior frontal gyrus (SFG) was significantly activated with increasing levels of envy. In fMRI studies on love–envy, Katrin et al. (2015) used an infidelity contextual utterance task to induce love–envy and found that the anterior cingulate cortex (ACC) was activated, while Sun et al. (2016) used a contextual imagery task to induce love–envy and found that the ventral medial prefrontal cortex (VMPFC) was activated.

Previous studies have some limitations. For instance, empirical research has its own inherent shortcomings. Firstly, individual brain imaging studies tend to involve a relatively small number of subjects. This may lead to low statistical test power and effect sizes (Yarkoni, 2009). Secondly, neuroimaging results may be inconsistent due to the sensitivity of the task and control conditions selected. Thirdly, Single fMRI studies often focus only on specific activated brain regions related to envy, disregarding the broader mechanisms responsible for generating and regulating it. Therefore, meta-analysis techniques based on large-scale data synthesis methods are necessary to overcome the limitations of individual brain imaging studies (Yarkoni et al., 2011). This method not only helps to make up for the lack of understanding of the three envy types as a whole, but also explores the generality and variability of neural activity among the three types, and provides representative reference coordinate points for future region of interest (ROI) analyses. However, there is a lack of meta-analytic studies comparing the trait (personality) envy, social comparison envy, and love–envy, and the understanding of the similarities and differences in the neural mechanisms behind them is relatively unclear.

Activation likelihood estimation (ALE) meta-analysis is an unbiased and objective approach to analyzing brain function (Wager et al., 2007). It can provide a consistent quantitative measure of relevant studies in this research area. Notably, the ALE meta-analysis method effectively avoids the problems of low statistical test power and high false-favorable rates in individual neuroimaging studies (Button et al., 2013; Eklund et al., 2016). Therefore, this study analyzed the existing functional magnetic resonance imaging (fMRI) studies of three envy types, namely trait (personality) envy, social comparison envy, and love–envy, by ALE and observed the similarities and differences in the processing brain regions of the three envy types to identify the neural mechanisms underlying the processing of the three envy types.

2 Methods

2.1 Literature search and inclusion criteria

This study used the CNKI full-text database to search for Chinese literature and PubMed, Web of Science, Elsevier Science Direct, Semantic Scholar, and ProQuest databases to search for foreign language literature. To conduct the literature search, we used the keywords "envy" and "fMRI" for Chinese sources and "envy," "fMRI," or "envy and fMRI" (adapted for the Web of Science database format) for foreign sources. After the screening, we obtained 425 papers. Further, after reading the abstract, methods, and results sections of each article, those that met the following six characteristics were included in the meta-analysis:

1 The type of study in question is empirical literature, which excludes reviews, meta-analyses, and case studies.
2 The research content excludes experienced envy, attributed envy, benign envy, malicious envy, and other types of envy less studied, focusing on trait (personality) envy, social comparison envy, and love envy.
3 The study only included normal individuals as subjects. Patients with brain lesions, neurological conditions, juvenile delinquents, and other special groups whose brain structure and function have been significantly altered were not part of the meta-analysis.
4 The research methodology involved subjects completing a scale or an experimental task related to jealousy, with an experimental comparison condition related to jealousy (contrast). The study used the fMRI method, excluding other methods like electroencephalography (EEG), magnetoencephalography (MEG), diffusion tensor imaging (DTI), which were used to analyze the condition of white and gray matter, and magnetic resonance imaging (MRI).
5 Whole-brain analyses were used, excluding studies with only region of interest analysis.
6 The study provided the coordinates of the brain regions that were found to be activated during the experiment. The

activation results were reported using the standardized Talairach or MNI space. We excluded studies that did not report the coordinates of the activated regions. We also removed peak coordinates that were unrelated to envy and only activated by the evoked task, such as the peak activation coordinates in the visual cortex.

A total of 13 papers finally met the above criteria and were included in the present study's meta-analysis. Figure 1 shows the specific screening process and results.

2.2 Systematic review

We followed recent recommendations on how to conduct a proper neuroimaging meta-analysis (Müller et al., 2018). For the current meta-analysis, 13 studies met the inclusion criteria reported in the previous section. Table 1 provides the basic information on the included studies.

Data were extracted from the studies and then checked. We then created a database containing the following information of the selected articles: type of envy, literature information (author and publication date), number of participants and among them the number of female participants, average age of participants, experimental comparison conditions, number of activation peaks reported for the experiments, and coordinate space (Talairach or MNI space).

The 13 included literature reported three types of envy, a total of 698 participants. In the literature on trait (personality) envy, there were 186 male participants and 192 female participants. In the literature on social comparison envy, there were 143 male participants and 88 female participants. In the literature on love–envy, there were 51 male participants and 38 female participants. The included literature had an average age ranging from 17 to 27 years. Their average age ranged from 17 to 27 years. The 13 included literature also reported 40 experimental comparison conditions and 216 peaks. For most functional neuroimaging meta-analyses, it is important to explicitly incorporate the paradigm of the literature (Müller et al., 2018). This paper considered all paradigms for different types of envy and focused on the higher order supervisory control processes necessary in all paradigm types. Of the 40 comparison conditions, trait (personality) envy included three approaches: regional homogeneity (ReHo), voxel-based morphometry (VBM), and neural representations of emotion. There are two main task types to induce social comparison envy - the story context method and the money gain/loss game; the story context method causes the SpHi condition (SpHi = superior with high similarity), the SpLo condition (SpLo = superior with low similarity), AvLo condition (AvLo = average with low similarity) three scenarios or target character and positive or unfortunate (fortunate/neutral) events, the money gain/loss game induces gain, loss, no change or Ro (reward for other) and Rs (reward for self) game outcomes. The infidelity contextual statement task comprises sexual infidelity, emotional infidelity, and neutral contextual

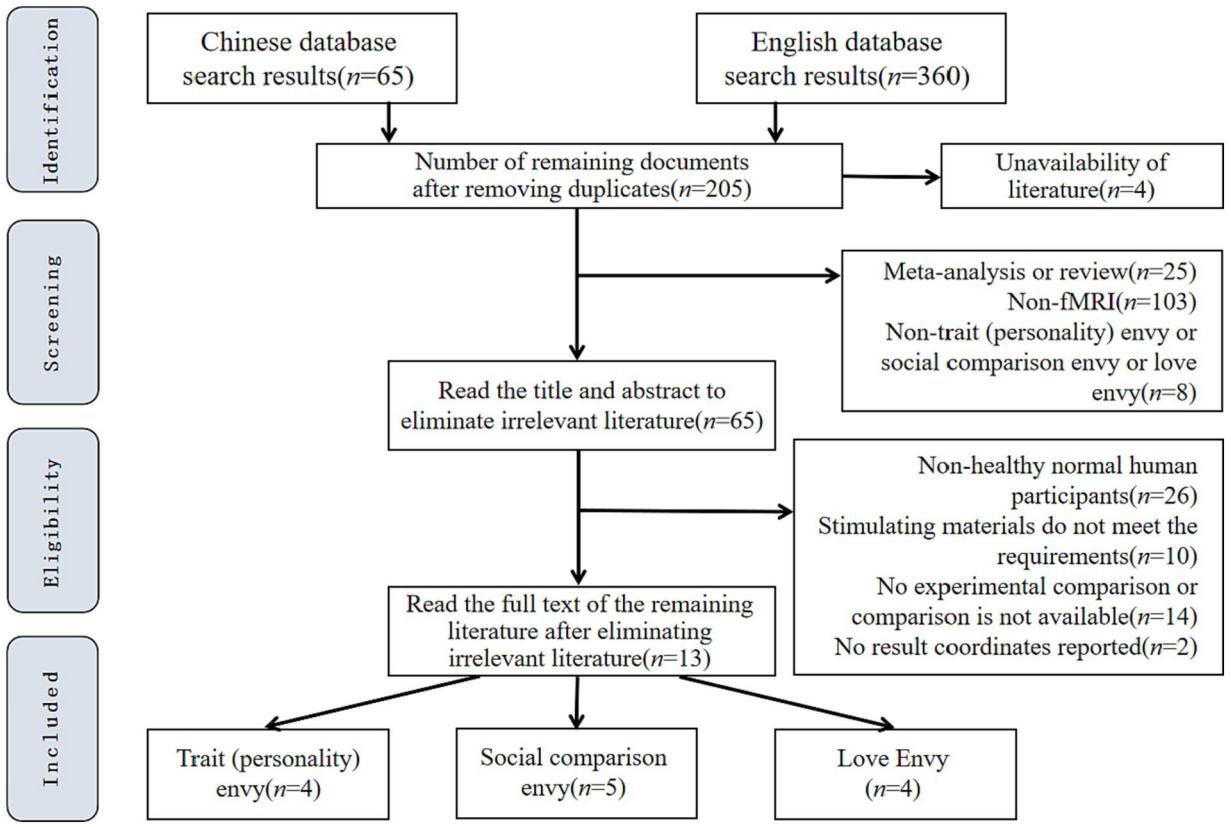


FIGURE 1
Flow chart of literature screening.

TABLE 1 Details of the 13 included literature.

Types of envy	Literature information	*N*, Sex	Age	Contrasts	Peak	Coordinate system
Trait (personality) envy	Zheng (2021)	218, 107F	21.42	2, Angry, baseline	20	MNI
	Zheng et al. (2019)	92, 45F	21,68	5, Angry, happy, fear, sad, neutral	21	MNI
	Xiang et al. (2017)	27, 16F	20.63	2, Dispositional envy, neutral	2	MNI
	Xiang et al. (2016)	41, 24F	21,37	2, Dispositional envy, neutral	2	MNI
Social comparison envy	Sol et al. (2023)	58, 27F	27.86	2, Envy, neutral	15	MNI
	Daniel et al. (2021)	39, 0F	17.16	3, Positive, negative, neutral outcomes	11	MNI
	Brennan et al. (2020)	19, 10F	27.2	3, SpHi (superior with high similarity) condition, SpLo (superior with low similarity) condition, AvLo (average with low similarity) condition	4	MNI
	Tanaka et al. (2019)	97, 41F	19.3	2, Ro (reward for other)、Rs (reward for self)	1	MNI
	Dvash et al. (2010)	18, 10F	26.76	5, The absolute gain events, the other's greater gain, the absolute loss events, the other's greater loss, no change	13	Talairach
love–envy	Nadine et al. (2019)	11, 11F	29.9	3, Jealousy Condition (JC), Control Condition (CC), Nonsense words (NC)	62	MNI
	Sun et al. (2016)	37, 18F	22.8	2, Happiness scenarios, Jealousy scenarios	15	MNI
	Katrin et al. (2015)	22, 0F	26.73	3, Sexual infidelity, Emotional infidelity, Neutral	23	MNI
	Takahashi et al. (2006)	19, 9F	22.1	6, Men (Sexual infidelity, Emotional infidelity, Neutral), Women (Sexual infidelity, Emotional infidelity, Neutral)	21	MNI

The MNI space is a coordinate system created by the Montreal Neurological Institute based on a series of magnetic resonance images of the average human brain. The Talairach coordinate system is based on the standard brain anatomy atlas established by French anatomist Talairach.

statements. The contextual imagination task involves imagining a familiar friend with the love rival "A," the love object "B," and the love rival "Jack." It includes two scenarios: happiness scenarios, where the participant imagines themselves with "A" or "B," envy scenarios, where the participant imagines "Jack" interacting with "A" or "B," and "A" or "B" interaction scenarios.

2.3 Publication bias

Publication bias is a problem that should be addressed. That is, there is in general in science a bias to publish mainly significant results while experiments failing to reject the null-hypothesis are often not reported (Ioannidis et al., 2014). Publication bias seriously impacts the reliability of meta-analysis results and overestimates the existing average effect. Several methods are generally used to test for publication bias meta-analyses, including the funnel plot, Begg test, classic fail-safe *N* test, Egger's test, and *p*-curve test. These tests can help determine whether there is significant publication bias in a meta-analysis. In this study, a funnel chart and Begg test were used to test for publication bias.

Note that the literature we included was all related to the study of various types of envy and fMRI. Since the data included in the meta-analysis were the fMRI coordinates used in various literature, publication bias could not be determined from this. Therefore, this paper uses the main effect size of each literature to judge the problem of publication bias.

2.4 Activation likelihood estimation method

ALE analysis is a meta-analytic technique that evaluates the co-localization of reported activations across studies. The first step is to categorize experiments in the literature, such as by stimulus or task. Whole-brain probability maps are then created across the reported foci in standardized stereotaxic space (Talairach or MNI). To create probability maps, this meta-analysis used GingerALE software. The probabilities are modeled by 3D Gaussian density distributions that adjust the FWHM for each study to account for sample size variability. For each voxel, GingerALE estimates the cumulative probabilities that at least one study reports activation for that locus. This generates a statistically thresholded ALE map, accounting for spatial uncertainty across reports. The resulting ALE values reflect the probability of reported activation at that locus, with high values indicating high probability estimates. This value is tested against the null hypothesis that activation is independently distributed across all studies in the meta-analysis, using random effects (Turkeltaub et al., 2012).

The GingerALE software (V3.0.2)[1] was used to process the data. However, there was a problem of inconsistency in the peak coordinate system as some studies used the Talairach coordinate system while others used the MNI (Montreal Neurological Institute) coordinate

1 http://brainmap.org/ale/

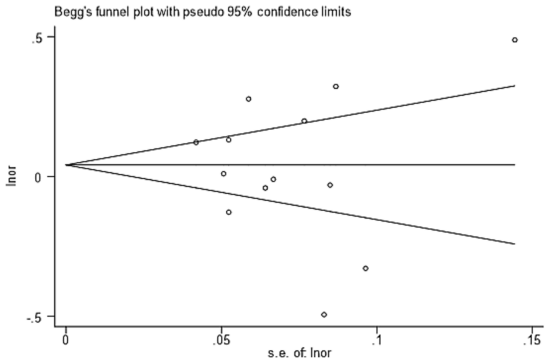


FIGURE 2
The result of funnel diagram.

system based on the standard brain template on which the peak coordinates were derived. To address this issue, before analyzing the data, the coordinate systems of the studies included in the analysis were converted using the Convert Foci tool in the GingerALE software. The "Brett: MNI to Talairach" option was selected to convert the reported coordinates from MNI space to Talairach space. This conversion is done automatically when the data are inserted into the BrianMap database using a transform called icbm2tal developed by Lancaster et al. (2007). This new transform provides improved fit over the Brett transform (mni2tal), and improves the accuracy of meta-analyses (Laird et al., 2010).

Afterward, the process of "Single Dataset" was carried out for the three types of envy. Talairach contrast coordinates of activation from eligible envy studies were combined (use the "Save and Merge Foci" tool in GingerALE) to create 3D maps depicting the likelihood of activation within each voxel in an fMRI template. Significant areas were identified depending on whether the envy processing location was more likely to occur in comparison to random spatial distributions. Analyses were thresholded using a cluster-lever FWE for multiple comparisons at $p = 0.05$. The FWE corrected threshold is set to the ALE value that no more than a specified fraction of the distribution exceeds that value. FWE thresholds are more conservative, so 5% of random studies, or $p < 0.05$ is recommended (Eickhoff et al., 2016). Finally, using multiple comparisons (5,000 alignments) correction at a clustering threshold of $p < 0.001$ (Liu et al., 2022).

Contrast analyses were performed to identify common (i.e., conjunction) and significantly different brain areas involved in trait (personality) envy, comparison envy, and love envy. Since the contrast analyses used ALE maps thresholded for multiple comparisons, the threshold was set to uncorrected $p = 0.01$ (10,000 permutations, 200 mm^2 minimum volume for contrasts; Arsalidou et al., 2020). With these options, GingerALE software allows for between group comparisons, however, currently there are no options for correlational analyses.

Once the thresholded map has been created, we'll need an anatomical underlay in order to view the meta-analysis results in context. Mango (Multi-Image Analysis GUI)[2] is a viewer for biomedical research images developed by Jack Lancaster and Michael Martinez. We use the "Colin_tlrc_ 2×2×2. Nii (dimensions match GingerALE images)" to view our meta-analysis results on Mango (V4.1).

2 http://ric.Uthscsa.edu/mango/

3 Results

3.1 Publication bias

Through the funnel diagram, it can be found that the effect distribution represented by 13 included literature is roughly symmetrical in the funnel plot, in which there are six points above the average effect value and seven points below the average effect value. Both above and below, there are only three points outside the 95% confidence region. Therefore, the funnel plot can show that there is no publication bias in the included literature research. The p value obtained by the Begg test is 0.760, which also confirms the above view. The result of funnel diagram is shown in Figure 2.

3.2 Single meta-analysis results

The single meta-analysis showed that 45 peak copulas for trait (personality) envy yielded 20 clusters, 44 peak copulas for social comparison envy yielded 25 clusters, and 121 peak copulas for love–envy yielded 45 clusters (Tables 2–4). Calculation of the peak distribution ratios for each region revealed that trait (personality) envy-related peaks were mainly distributed in the frontal lobe (35%), parietal lobe (15%), posterior lobe (12%), Sub-lobar (12%), limbic lobe (9%), temporal lobe (8%), occipital lobe (5%), and anterior lobe (4%); social comparison envy-related peaks were mainly distributed in the frontal lobe (48%), parietal lobe (15%), temporal lobe (9%), posterior lobe (8%), limbic lobe (8%), occipital lobe (6%), sub-lobar (4%), anterior lobe (2%); love–envy-related peaks were mainly distributed in the frontal lobe (30%), sub-lobar (22%), limbic lobe (17%), parietal lobe (12%), temporal lobe (9%), occipital lobe (4%), posterior lobe (3%), anterior lobe (3%). Specifically, trait (personality) envy was activated mainly in the inferior frontal gyrus, cingulate gyrus, middle frontal gyrus, lentiform nucleus, inferior parietal lobule, declive, and superior frontal gyrus; social comparison envy was activated mainly in the middle frontal gyrus, inferior frontal gyrus, medial frontal gyrus, precuneus, inferior parietal lobule, precentral gyrus, superior temporal gyrus, declive, and anterior cingulate gyrus; love–envy was activated mainly in the inferior frontal gyrus, superior frontal gyrus, cingulate gyrus, insula, claustrum, medial frontal gyrus, inferior parietal lobule, caudate, and posterior cingulate gyrus. The distribution of brain activation in the three envy types is shown in Figure 3.

3.3 Joint analysis results

The results of the joint analysis showed that the fusion of trait (personality) envy and social comparison envy yielded a total of eight clusters, the fusion of trait (personality) envy and love–envy yielded a total of eight clusters, and social comparison envy and love–envy yielded a total of 10 clusters (Table 5). Specifically, trait (personality)

10.3389/fpsyg.2024.1335548

TABLE 2 Single meta-analysis of trait (personality) envy.

Cluster#	Volume mm^3	Hemisphere	Brain region	Peak coordinates (MNI coordinate system)			ALE Value ($\times 10^{-4}$)
				X	Y	Z	
1	7,368	R	FL (IFG)	30	18	−14	9.52
		R	Sub-lobar	36	4	−8	9.52
		R	FL (IFG)	44	16	−6	9.51
		R	FL (IFG)	32	30	−4	9.44
2	7,224	L	LL (AC)	−12	24	26	9.54
		L	FL (CC)	−12	22	36	9.49
		R	LL (CC)	2	22	38	9.47
		L	FL (SFG)	−20	20	48	9.45
3	5,888	L	TL (SFG)	−50	4	2	9.52
		L	FL (Prec)	−44	14	6	9.50
		L	TL (MTG)	−54	2	−8	9.50
		L	FL (Prec)	−48	16	6	9.47
4	5,520	L	Sub-lobar (LN)	−24	−18	10	9.48
		L	Sub-lobar (LN)	−30	−6	4	9.47
		L	Sub-lobar (LN)	−20	0	12	9.45
5	5,360	R	PL (IPL)	44	−34	36	9.51
		R	PL (Prec)	54	−28	44	9.51
		R	PL (IPL)	44	−48	38	9.48
6	4,960	\	\	24	24	32	9.50
		R	FL (Sub-lobar)	20	22	40	9.48
		R	FL (MFG)	32	26	26	9.48
7	4,952	L	PL (Precuneus)	−24	−64	38	9.49
		L	PL (Precuneus)	−14	−56	54	9.49
		L	PL (Precuneus)	−20	−62	46	9.48
8	4,648	R	FL (MFG)	48	30	16	9.49
		R	FL (MFG)	50	30	22	9.49
		R	FL (MFG)	50	16	28	9.44
9	4,488	L	FL (IFG)	−44	18	−12	9.49
		L	FL (IFG)	−26	22	−10	9.49
		L	FL (IFG)	−38	18	−10	9.43
10	3,712	R	Sub-lobar (LN)	20	−12	0	9.48
		R	Sub-lobar (LN)	32	−18	−8	9.46
11	3,656	L	PL (IPL)	−54	−34	40	9.51
		L	PL (IPL)	−42	−42	38	9.44
12	3,328	L	FL (ParL)	−2	−30	48	9.45
		L	FL (ParL)	−2	−28	56	9.43
13	1872	R	PL	38	−64	−24	9.48
14	1856	L	LL (CG)	−8	−26	34	9.45
15	1848	L	FL (SFG)	−18	48	20	9.48
16	1840	L	PoL	−6	−70	−24	9.46
17	1840	L	PoL	−12	−58	−14	9.48
18	1832	R	TL (FG)	36	−42	−14	9.52
19	1824	R	OL (IOG)	44	−72	−6	9.46
20	1824	R	FL (IFG)	48	14	12	9.54

FL, frontal lobe; IFG, inferior frontal gyrus; LL, limbic lobe; AC, anterior cingulate; CG, cingulate gyrus; SFG, superior frontal gyrus; TL, temporal lobe; Prec., precentral gyrus; ParL, paracentral lobule; MTG, middle temporal gyrus; LN, lentiform nucleus; PL, parietal lobe; IPL, inferior parietal lobule; MFG, middle frontal gyrus; PoL, posterior lobe; FG, fusiform gyrus; OL, occipital lobe; IOG, inferior occipital gyrus.

10.3389/fpsyg.2024.1335548

TABLE 3 Single meta-analysis of social comparison envy.

Cluster#	Volume mm^3	Hemisphere	Brain region	Peak coordinates (MNI coordinate system)			ALE Value ($\times 10^{-4}$)
				X	Y	Z	
1	9,464	R	PL (Precuneus)	12	−64	48	9.54
		R	PL (Precuneus)	12	−56	44	9.51
		L	PL (Precuneus)	−2	−58	30	9.50
		R	PL (Precuneus)	8	−50	40	9.49
		R	PL (Precuneus)	8	−42	46	9.46
		R	PL (Precuneus)	6	−62	38	9.45
2	7,656	L	FL (SFG)	−18	42	26	9.55
		R	LL (CG)	2	25	32	9.49
		L	FL (MFG)	−30	34	34	9.48
		L	LL (AG)	−2	36	24	9.42
3	5,792	L	TL (STG)	−38	8	−30	9.49
		L	LL	−28	6	−20	9.48
		L	LL (PG)	−30	−6	−14	9.47
4	5,176	R	FL (Prec)	52	12	8	9.51
		R	FL (IFG)	52	6	14	9.50
		R	FL (IFG)	48	20	0	9.46
5	3,824	R	PL (PoG)	54	−28	40	9.47
		R	PL (SG)	56	−36	32	9.45
6	3,720	L	LL (AC)	−8	42	8	9.50
		R	LL (AC)	5	44	10	9.43
7	3,704	R	FL (MFG)	38	40	22	9.55
		R	FL (MFG)	30	34	28	9.53
8	3,440	R	FL (MeFG)	2	28	−14	9.55
		L	FL (MeFG)	−2	22	−16	9.43
9	3,392	L	TL (ITG)	−50	−52	−8	9.54
		L	TL (MTF)	−54	−44	−10	9.50
10	2,672	L	FL (IFG)	−40	20	−16	9.51
		L	TL (STG)	−32	24	−24	9.47
11	2,376	L	PL (Precuneus)	−12	−56	44	9.51
12	1984	R	OL (FG)	23	−58	−8	9.40
13	1952	L	PoL	−22	−58	−12	9.55
14	1936	R	Sub-lobar	30	18	6	9.50
15	1928	R	PL (IPL)	46	−50	40	9.49
16	1920	R	OL (LG)	4	−86	0	9.47
17	1912	R	FL (MeFG)	2	−14	64	9.51
18	1904	L	FL (Prec)	−42	4	37	9.49
19	1896	R	FL (IFG)	39	20	−18	9.48
20	1880	R	FL (MFG)	42	2	38	9.50
21	1,608	R	PoL	46	−68	−30	9.47
22	1,600	R	FL (SFG)	30	54	−4	9.50
23	1,504	L	FL (SFG)	−10	42	48	9.53
24	1,120	L	FL (SFG)	−17	62	24	9.44
25	864	R	FL (SFG)	12	66	18	9.48

PL, parietal lobe; FL, frontal lobe; LL, limbic lobe; SFG, superior frontal gyrus; CG, cingulate gyrus; MFG, middle frontal gyrus; AC, anterior cingulate; MeFG, medial frontal gyrus; TL, temporal lobe; STG, superior temporal gyrus; PG, parahippocampal gyrus; Prec, precentral gyrus; IFG, inferior frontal gyrus; PoG, postcentral gyrus; SG, supramarginal gyrus; ITG, inferior temporal gyrus; MTG, middle temporal gyrus; OL, occipital lobe; PoL, posterior lobe; FG, fusiform gyrus; IPL, inferior parietal lobule; LG, lingual gyrus.

10.3389/fpsyg.2024.1335548

TABLE 4 Single meta-analysis of love–envy.

Cluster#	Volume mm^3	Hemisphere	Brain region	Peak coordinates (MNI coordinate system)			ALE Value (×10^{-4})
				X	Y	Z	
1	21,440	L	Sub-lobar	−8	2	0	9.56
		L	Midbrain	0	−14	−8	9.56
		R	Midbrain	12	−12	−2	9.56
		L	Sub-lobar	−18	−8	−2	9.55
		L	Sub-lobar	−12	2	2	9.55
		L	Sub-lobar	−24	−10	−2	9.55
		L	Sub-lobar	−12	12	2	9.55
		R	FL (IFG)	24	8	−18	9.55
		L	Midbrain	−6	−18	−2	9.54
		R	Midbrain	8	−18	−2	9.54
		L	Midbrain	−6	−12	−4	9.54
		L	Midbrain	−8	−14	−10	9.54
		R	Sub-lobar	12	0	−2	9.54
		R	Sub-lobar	6	−2	6	9.52
		L	Sub-lobar	−32	−28	−2	9.52
		R	LL	22	2	−12	9.51
		L	Midbrain	−16	−24	−2	9.51
		L	Sub-lobar	−34	−26	−6	9.51
		L	Sub-lobar	−12	−6	10	9.51
		L	TL	−30	−20	−10	9.51
		L	Sub-lobar	−8	−12	8	9.50
		R	Sub-lobar	18	−4	−2	9.48
		L	Sub-lobar	−6	10	−2	9.48
		L	LL	−24	−20	−6	9.48
		L	Sub-lobar	−18	8	−2	9.48
		R	Sub-lobar	2	−4	−10	9.47
		L	Sub-lobar	−20	−24	−2	9.47
		L	Sub-lobar	−6	−30	2	9.47
		L	Midbrain	−8	−18	−10	9.47
		L	Sub-lobar	−4	−24	6	9.46
		R	Midbrain	8	−12	−8	9.43
		L	Sub-lobar	−8	−6	12	9.42
		R	Sub-lobar	2	−10	8	9.41
		L	Sub-lobar	−2	−4	−10	9.41
		L	Sub-lobar	−14	12	6	9.40
2	4,096	R	PoL	20	−66	−16	9.54
		R	PoL	12	−74	−8	9.53
		R	PoL	10	−64	−22	9.52
		R	PoL	18	−68	−12	9.49
		R	PoL	24	−86	−19	9.47
		R	PoL	16	−80	−16	9.47
3	3,520	L	LL (CG)	−2	36	28	9.54

(Continued)

10.3389/fpsyg.2024.1335548

TABLE 4 (Continued)

Cluster#	Volume mm^3	Hemisphere	Brain region	Peak coordinates (MNI coordinate system)			ALE Value (×10^{-4})
				X	Y	Z	
		L	LL (AC)	−4	26	24	9.52
		L	LL (AC)	−2	20	20	9.51
		R	FL (MeFG)	8	42	26	9.49
		L	FL (MeFG)	−2	44	22	9.39
4	2,584	R	LL (AC)	4	36	6	9.51
		L	LL (AC)	−2	36	−4	9.47
		R	LL (AC)	2	32	12	9.45
		L	LL (AC)	−2	32	14	9.40
5	2,520	R	TL (SG)	64	−46	24	9.49
		R	TL (STG)	50	−46	16	9.48
		R	TL (STG)	62	−48	16	9.48
		R	TL (STG)	60	−58	22	9.43
6	2,312	R	Sub-lobar	18	12	−6	9.51
		R	Sub-lobar	14	12	8	9.50
		R	Sub-lobar	8	12	4	9.43
7	2,224	R	LL (PG)	28	−22	−14	9.55
		R	Sub-lobar	34	−26	−6	9.51
		R	LL (PG)	24	−20	−6	9.48
8	1872	L	TL (MTG)	−46	−76	28	9.53
		L	TL (MTG)	−52	−70	22	9.50
		L	TL (MTG)	−40	−66	22	9.48
9	1848	L	PL (IPL)	−50	−50	38	9.52
		L	PL (IPL)	−50	−42	34	9.51
		L	PL (IPL)	−60	−46	40	9.45
10	1,648	R	LL (CG)	6	−18	34	9.53
		L	LL (CG)	0	−16	38	9.53
11	1,464	L	FL (IFG)	−38	14	−10	9.50
		L	Sub-lobar	−30	12	−8	9.47
12	1,456	L	FL (IFG)	−38	28	−2	9.47
		L	FL (IFG)	−44	28	6	9.47
13	1,456	L	PL (Precuneus)	−8	−60	22	9.50
		L	LL (PoC)	−8	−52	16	9.48
14	1,456	R	FL (SFG)	6	44	44	9.50
		R	FL (MeFG)	12	38	38	9.46
15	1,448	L	FL (Prec)	−14	−30	64	9.45
		L	FL (MeFG)	−2	−26	62	9.38
16	1,440	R	PL (IPL)	50	−36	32	9.54
		R	PL (SG)	40	−38	32	9.49
17	1,416	L	FL (MeFG)	−8	50	10	9.48
		L	FL (MeFG)	−4	50	4	9.47
18	1,400	L	TL (MTG)	−56	−48	6	9.53
		L	TL (STG)	−60	−52	16	9.49

(Continued)

10.3389/fpsyg.2024.1335548

TABLE 4 (Continued)

Cluster#	Volume mm^3	Hemisphere	Brain region	Peak coordinates (MNI coordinate system)			ALE Value (×10^{-4})
				X	Y	Z	
19	1,344	L	FL (SFG)	−2	24	52	9.55
		L	FL (SFG)	−6	20	60	9.51
20	1,280	R	LL (PoC)	6	−54	22	9.56
		R	LL (PoC)	8	−52	16	9.48
21	880	L	FL (MFG)	−44	8	46	9.53
		L	FL (MFG)	−48	6	46	9.53
22	752	R	AL	6	−38	−4	9.48
23	752	L	Sub-lobar	−34	3	2	9.47
24	752	L	Sub-lobar	−38	18	12	9.46
25	736	L	PoL	−24	−86	−19	9.47
26	736	L	OL (MTG)	−44	−76	12	9.44
27	736	L	PL (SG)	−32	−50	32	9.44
28	736	L	LL (CG)	−12	−42	36	9.50
29	736	L	PL (PoC)	−30	−26	40	9.44
30	728	R	Sub-lobar	12	2	16	9.50
31	720	L	LL (PG)	−38	−48	−6	9.48
32	720	L	OL	−20	−70	12	9.47
33	720	R	Sub-lobar	18	−12	12	9.48
34	720	R	Sub-lobar	36	−34	20	9.48
35	720	L	LL (CG)	−18	−8	36	9.53
36	712	L	Sub-lobar	−44	14	0	9.52
37	712	L	FL (Prec)	−42	4	10	9.48
38	712	R	Sub-lobar	26	18	10	9.52
39	704	R	FL (IFG)	42	26	4	9.52
40	704	L	Sub-lobar	−40	−6	44	9.51
41	696	L	FL (AG)	−44	−62	34	9.53
42	688	L	LL (CG)	−18	−34	44	9.55
43	656	R	FL (SFG)	8	58	28	9.43
44	560	R	FL (IFG)	54	28	2	9.52
45	536	L	FL (MeFG)	−6	64	12	9.55

FL, frontal lobe; LL, limbic lobe; IFG, inferior frontal gyrus; PG, parahippocampal gyrus; TL, temporal lobe; PoL, posterior lobe; CG, cingulate gyrus; AC, anterior cingulate; MeFG, medial frontal gyrus; MFG, middle frontal gyrus; SG, supramarginal gyrus; STG, superior temporal gyrus; PL, parietal lobe; MTG, middle temporal gyrus; IPL, inferior parietal lobule; AL, anterior lobe; Precuneus; PoC, posterior cingulate; SFG, superior frontal gyrus; Prec, precentral gyrus; OL, occipital lobe; PoG, postcentral gyrus; AG, angular gyrus.

envy and social comparison envy co-activate the following brain regions: right frontal sub-Gyral, right inferior parietal lobule, left inferior frontal gyrus, left precuneus, right paracentral lobule, left posterior lobule declive, right posterior lobule, left extra-nuclear lobule; trait (personality) envy and love–envy co-activate the following brain regions: left lobule extra the brain areas co-activated by trait envy and love–envy are: left extra-nuclear lobule, right sub-lobar lentiform nucleus, left paracentral lobule, left parietal supramarginal gyrus, left limbic cingulate gyrus, right inferior frontal gyrus, right parietal supramarginal gyrus, left middle frontal gyrus; brain areas co-activated by social comparison envy and love–envy are: left limbic anterior cingulate gyrus, right sub-lobar insula, right parietal supramarginal gyrus, left inferior frontal gyrus, cingulate gyrus, declive, middle frontal gyrus, temporal lobe sub-Gyral, and extra-nuclear lobule. The distribution of the activated brain areas jointly activated the brain between the two envy types is shown in Figure 4.

3.4 Contrasting analysis results

The results of the comparative analysis showed no activation clusters in the two-by-two comparative analysis of three envy types.

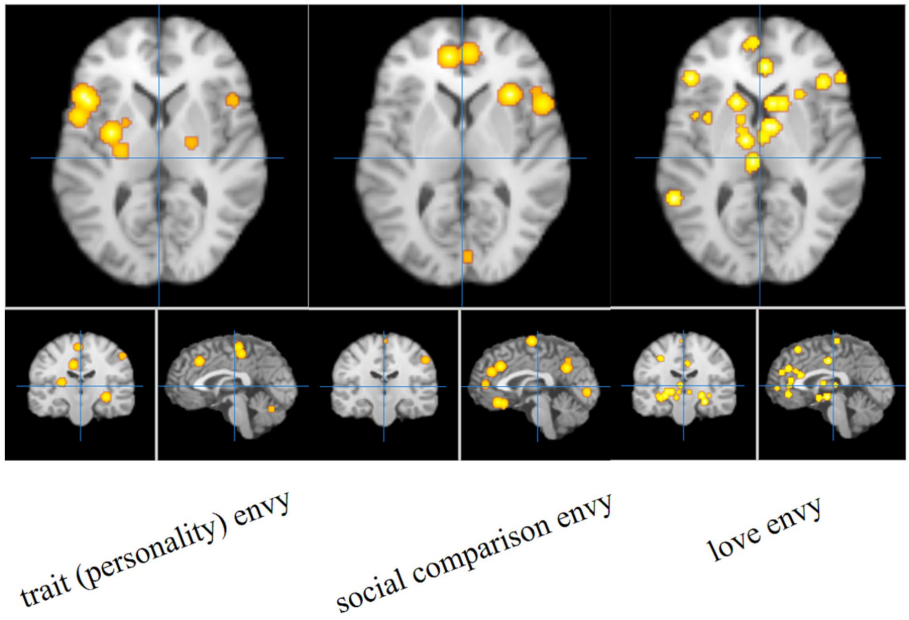


FIGURE 3
Single meta-analysis of the three envy types.

To reduce the possibility of type II statistical errors, ALE comparative analysis was also validated using very loose thresholds (no correction threshold, $p < 0.05$, and a minimum activation cluster size of 100 mm^3), and no significant differences were found.

4 Discussion

This study analyzed 13 existing functional magnetic resonance imaging studies of three types of envy – trait (personality) envy, social comparison envy, and love–envy – to observe the similarities and differences in the brain regions involved in the three types of envy processing and identify the neural mechanisms underlying the processing of the three types of envy.

4.1 Single-unit analysis: neural mechanisms of different types of envy

The single-unit analysis showed that 45 peak copula classes of trait (personality) envy yielded 20 clusters. The key brain regions that were activated included the inferior frontal gyrus, cingulate gyrus, middle frontal gyrus, lentiform nucleus, inferior parietal lobule, declive, and superior frontal gyrus.

Ochsner and Gross (2007) constructed a cognitive control model of emotion regulation with a bottom-up and top-down perspective. The brain neural network uses a bottom-up approach to encode the dynamic properties of stimuli, evaluate different types of emotions, and generate different types of emotional responses; it performs the evaluation of emotional stimuli and the control of emotional expression or experience in a top-down manner, regulating them, channeling them, and changing how emotional stimuli are evaluated.

Our account of how envy is generated is multi-leveled and bottom-up in its description of both the processes and the neural systems that give rise to emotional response. In the first step, a stimulus is perceived in its current situational context. The lentiform nucleus is the core region of the vertebrate neural circuit," and its activity is enhanced when exposed to negative emotional stimuli (Deng et al., 2015). In the included literature, researchers used the Multidimensional Jealousy Scale (MJS) and the Dispositional envy scale (DES) to measure participants' levels of trait envy. Example of statements include, "I feel envy every day," and "Feelings of envy constantly torment me." During the measurement, participants were asked to recall instances of envy. This led to activation of the lentiform nucleus. At the second stage, one attends to some of these stimuli or their attributes and appraises the significance of stimuli in terms of their relevance to one's current goals, wants or needs. The cingulate gyrus is located in the "core limbic brain cluster," a central structure responsible for integrating emotions, responding to emotional information and encoding information about emotionally salient events (Moraweta et al., 2017). Therefore, it is associated with the trait (personality) envy. When recalling instances of envy, the subjects emotionally encoded them. Finally, the third stage involves translating these appraisals into changes in experience, emotion-expressive behavior, and autonomic physiology.

With an understanding of how emotions are generated in the first place we can turn to an account of the process and neural systems involved in regulating them. Emotional regulation is top-down. The superior frontal gyrus and inferior parietal lobule are thought to be closely related to cognitive control and emotion regulation (Sheline et al., 2010; Thomas and Joseph, 2016). The generation of irrational envy accompanies the activation of both brain regions, and the levels of cognitive control and emotion regulation ability of different individuals affect their traits (personality). Studies have found that the middle and inferior

10.3389/fpsyg.2024.1335548

TABLE 5 Joint analysis of activation cluster results.

Conjunction	Volume mm^3	Hemisphere	Brain region	Center coordinates			ALE Value (×10^{-3})
				X	Y	Z	
trait_AND_social	47,976	R	FL (sub-Gyral)	23.4	25.5	15.4	18
	12,160	R	PL (IPL)	48.9	−37.9	38.4	19
	8,248	L	FL (IFG)	−36	18.2	−13.9	18
	6,968	L	PL (Precuneus)	−15.9	−58.8	46.3	16
	6,896	R	FL (ParL)	2.7	−33.6	51.5	12
	3,912	L	PoL (Declive)	−16.9	−57.7	−12.9	15
	3,800	R	PoL	41.7	−65.4	−26.1	16
	1,592	L	Sub-lobar (EN)	−29.7	−6.6	−4.9	10
Trait_AND_love	33,216	L	Sub-lobar (EN)	−32.4	5.7	1.7	18
	18,832	R	Sub-lobar (LN)	24	−9.8	−6.1	17
	10,672	L	FL (ParL)	−6.3	−26.9	47.1	17
	8,664	L	PL (SG)	−44.3	−45	36.2	16
	8,624	L	LL (CG)	−4.1	25.3	30.4	16
	7,792	R	FL (IFG)	44.2	25	4.7	14
	6,928	R	PL (SG)	45.2	−36.7	34.7	17
	2,840	L	FL (MFG)	−11.9	48.1	17.2	13
Social_AND_love	23,680	L	LL (AC)	−1.8	36.6	13.6	19
	9,456	R	Sub-lobar (Insula)	36.9	19.8	4.9	18
	7,376	R	PL (SG)	51.7	−37.4	32.8	18
	6,864	L	FL (IFG)	−36.2	15.3	−12.4	17
	5,496	L	LL (CG)	−1.3	−57.3	26.1	16
	4,800	R	PoL (Declive)	16.1	−69.1	−8.9	14
	4,488	L	FL (MFG)	−42.4	3.3	41.4	16
	4,280	L	TL (sub-Gyral)	−48.2	−49.2	−4.9	14
	3,712	L	Sub-lobar (EN)	−27.1	−11.3	−9.2	13

FL, frontal lobe; PL, parietal lobe; IPL, inferior parietal lobule; IFG, inferior frontal gyrus; Precuneus; ParL, paracentral lobule; PoL, posterior lobe; EN, extra-nuclear; LN, lentiform nucleus; SG, supramarginal gyrus; LL, limbic lobe; CG, cingulate gyrus; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; AC, anterior cingulate; TL, temporal lobe.

frontal gyrus are crucial for regulating negative emotions through cognitive reappraisal strategies and expression inhibition (Buhle et al., 2014; Nimarko et al., 2019). The core operation of expression inhibition is the individual's effort to inhibit emotion-related facial expressions when the stimulus has successfully evoked emotional expressions and physiological responses, such as breathing and heartbeat, to prevent emotions from being expressed when incentives have successfully produced them. The declive, as part of the cerebellar functional area, is involved in regulating muscle tone and coordinating the accuracy of casual movements (Dong et al., 2010) and may be related to the expression inhibition activity of facial expressions associated with trait (personality) envy.

In summary, we can classify the neural mechanisms of trait envy into three levels: "perception of negative stimuli," "encoding of emotional information," and "cognitive control and emotion regulation" (see Figure 5).

The single-unit analysis showed that the social comparison envy of the 44 peak copolymer classes yielded 25 clusters. The key brain regions primarily activated were the middle frontal gyrus, inferior frontal gyrus, medial frontal gyrus, precuneus, inferior parietal lobule, precentral gyrus, superior temporal gyrus, declive, and anterior cingulate gyrus.

We can find that the model of the neural mechanisms of social comparison envy remains consistent with the cognitive control model of emotion regulation (Ochsner and Gross, 2007). Regarding the process through which envy generates, in the first step, a stimulus is perceived in its current situational context. The generation of social comparative envy is closely linked to the social comparative context. Previous research has shown that the precuneus acquires information and experiences from a first-person perspective in highly integrated tasks while participating in contextual memory extraction, self-reference, and social cognition (Cavanna and Trimble, 2006; Buckner and Carroll, 2007). Contextual memory extraction is associated with the story-context approach for inducing social comparison envy. In contrast, self-reference and social cognition fit the defining characteristics of social comparison envy based on the upward social comparison that triggers envy. The medial frontal gyrus includes the medial prefrontal cortical and orbitofrontal regions and is associated with cognitive functions such as purposeful decision-making, and reward and punishment reflexes (Zhu, 2002). The functional magnetic

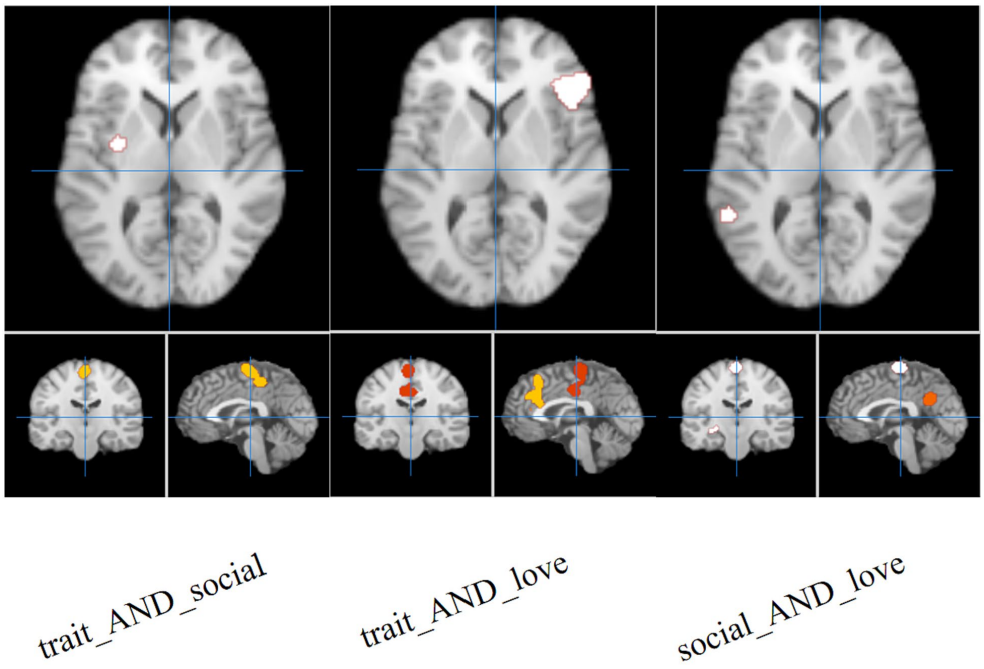


FIGURE 4
Joint analysis results. The intensity of activated brain regions in the figure gradually increases from red to white.

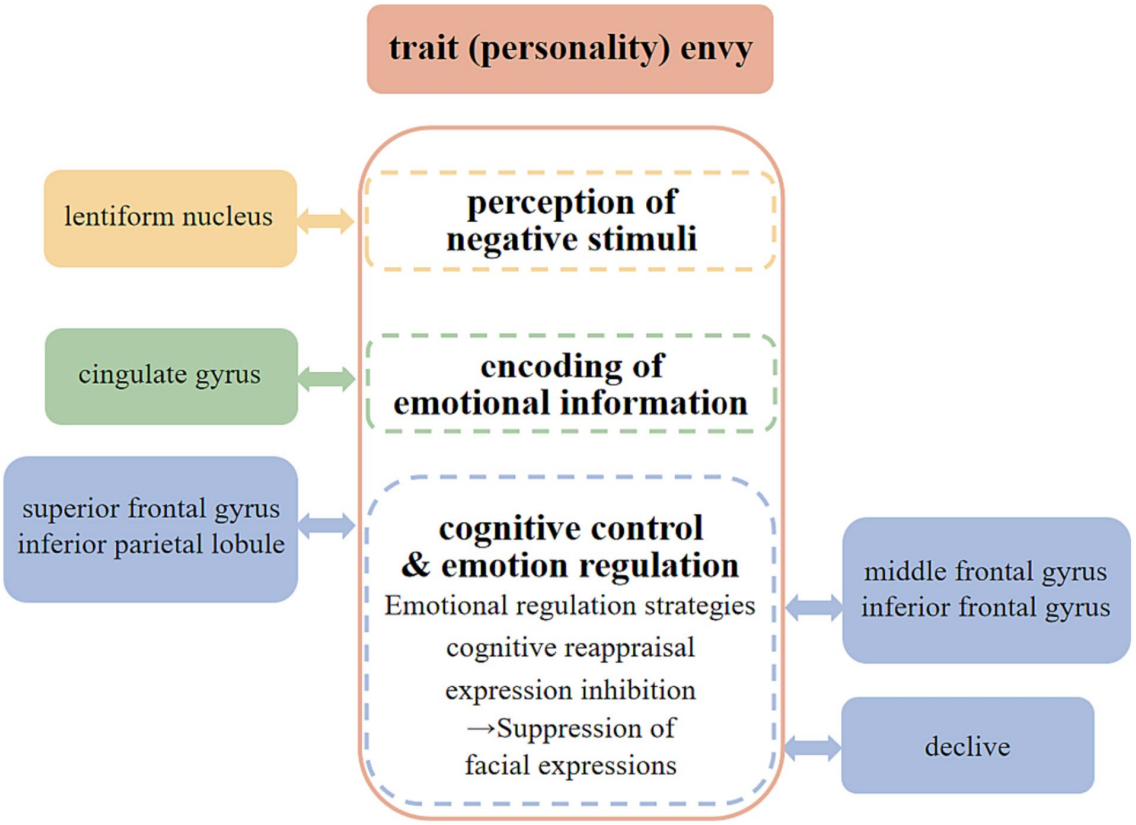


FIGURE 5
Neural mechanisms of the trait (personality) envy.

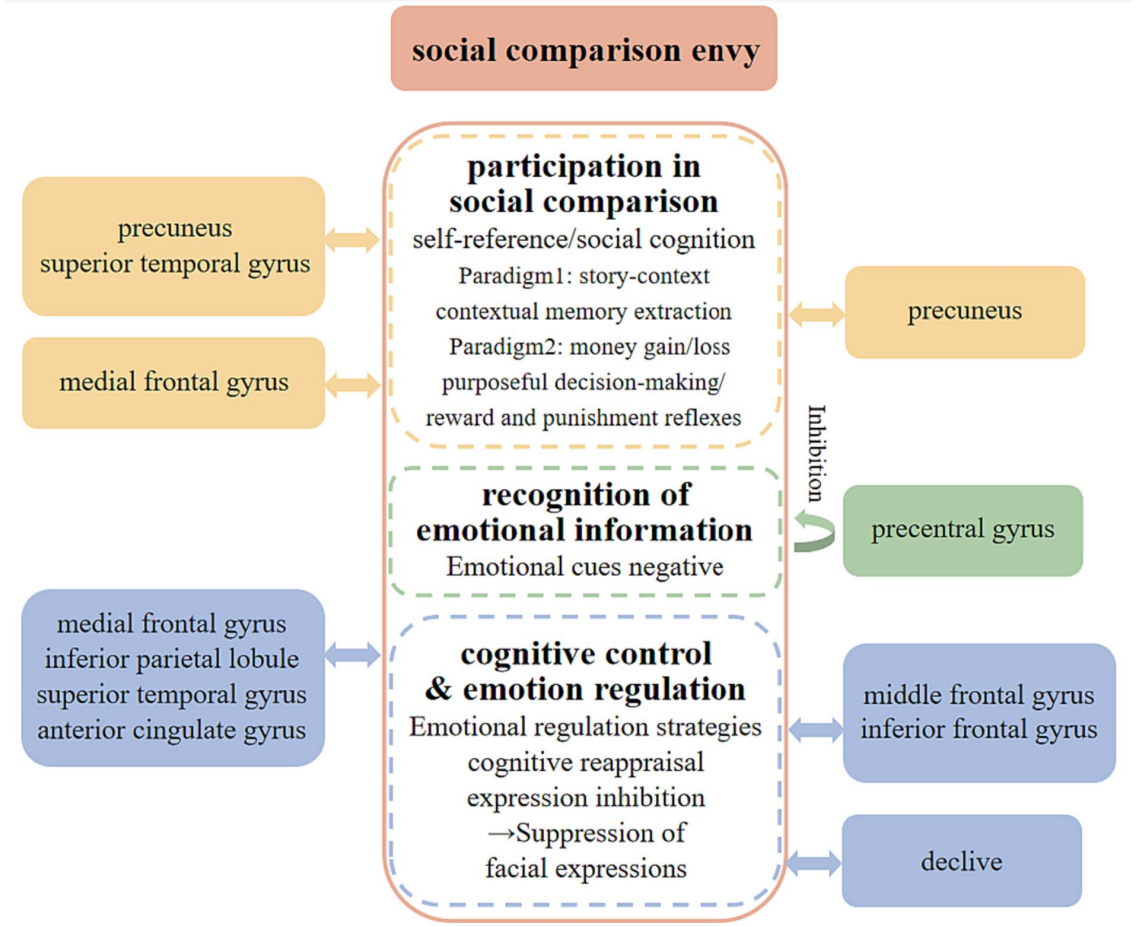


FIGURE 6
Neural mechanisms of social comparison envy.

resonance imaging study included in this study used the money gain/loss game to induce social comparison envy. The experimental paradigm begins with cognitive functions related to purposeful decision-making and the reward or punishment reflex. At the second stage, one attends to some of these stimuli or their attributes and appraises the significance of stimuli in terms of their relevance to one's current goals, wants or needs. The activation of the precentral gyrus is associated with oxytocin secretion (Chen, 2017), which influences individual emotion recognition. Olff et al. (2013) found that oxytocin enhances individual sensitivity to salience cues from the (social) environment (Bartz et al., 2011) or interpersonal (Ellenbogen et al., 2012), with the effect when salience cues are interpreted as "insecure" or "negative," oxytocin may inhibit the recognition of negative emotions from promoting socially adaptive behavior. Salient cues in situations that can provoke social comparison envy are insecure and negative for the individual. It appears that the generation of individual social comparison envy accompanies the activation of the precentral gyrus. Finally, the third stage involves translating these appraisals into changes in experience, emotion-expressive behavior, and autonomic physiology.

Regarding the process through which envy regulation, similar to the trait (personality) envy, envy regulation is top-down. Social comparison envy activates the inferior parietal lobule, which is closely associated with cognitive control and emotional regulation. The superior temporal gyrus plays a very important role in emotion regulation and social cognitive processing (Song et al., 2019). The medial frontal gyrus is also associated with emotion regulation (Zhu, 2002). The activated anterior cingulate gyrus can integrate afferent information from different sources and regulates cognitive and emotional functions (Bush et al., 2000). The middle and inferior frontal gyri exert cognitive reappraisal and expression-suppression strategies in emotion regulation. The expression-suppression strategy triggers the declive to engage in expression-suppression activities of facial expressions associated with social comparison envy.

In summary, we can classify the neural mechanisms of social comparison envy into three levels: "participation in social comparison," "recognition of emotional information," and "cognitive control and emotion regulation" (see Figure 6).

The single-unit analysis showed that 121 peak co-localization classes of love–envy yielded 45 clusters. The critical brain regions mainly activated include the inferior frontal gyrus, superior frontal gyrus, cingulate gyrus, insula, claustrum, medial frontal gyrus, inferior parietal lobule, caudate, and posterior cingulate gyrus.

We can find that the model of the neural mechanisms of social comparison envy remains consistent with the cognitive control model of emotion regulation (Ochsner and Gross, 2007). Regarding the

process through which envy generates, in the first step, a stimulus is perceived in its current situational context. The generation of love–envy is closely linked to context related to sexual infidelity and emotional infidelity. The claustrum plays a vital role in human sexual arousal and problems (Redouté et al., 2000), which fit the defining characteristics of love–envy based on sexual partnerships. The present study found that love–envy activated the caudate. The caudate involves reward mechanisms, emotional processing, and motivation (Villablanca, 2010). Neuroimaging studies of romantic love have found significantly more robust functional connectivity in the reward-motivation network (caudate) in relationship groups than in single groups (Song et al., 2015), suggesting that the triggering of love–envy is associated with the activation of the reward-motivation network by romantic love. The activation of the medial frontal gyrus is also associated with cognitive functions such as reward and punishment responses (Zhu, 2002). At the second stage, one attends to some of these stimuli or their attributes and appraises the significance of stimuli in terms of their relevance to one's current goals, wants or needs. The posterior cingulate gyrus is an assessment area (Bush et al., 2000) that integrates visual cognition in the visual cortex and emotional processes in the anterior cingulate gyrus in response to dynamic events (Lim et al., 2004; Enatsu et al., 2014; Morawetz et al., 2017). The love–envy brain imaging study selected for this study used an infidelity contextual utterance task. Peak visual cortical coordinates unrelated to envy and activated only by the evoked task were found with peak activation (removed). Similarly, posterior cingulate activation in visual-cortical visual-cognitive integration is task-related. Love–envy is induced by infidelity (sexual or emotional affairs) and dynamic events that activate the posterior cingulate gyrus. Therefore, love–envy is associated with the posterior cingulate gyrus. Finally, the third stage involves translating these appraisals into changes in experience, emotion-expressive behavior, and autonomic physiology. The insula is thought to represent a viscerotopic map of ascending viscerosensory inputs from the body (Mufson and Mesulam, 1982) and has been implicated in negative affective experience in general (Craig, 2009). There appears to be implicated in negative affective in the insula with posterior regions associated with primary representations of sensations from the body and anterior regions associated interoceptive awareness of the body and in motivational and affective states, like envy, that have a strong visceral component (Craig, 2009).

Regarding the process through which envy regulation, similar to the trait (personality) envy and social comparison envy, envy regulation is top-down. The insula is involved in various tasks related to emotional regulation and cognitive control (Cauda et al., 2012). Similar to the first two types of envy, love–envy activates the superior frontal gyrus, the medial frontal gyrus, and inferior parietal lobule, which are closely related to cognitive control and emotion regulation. The inferior frontal gyri exert cognitive reappraisal and expression-suppression strategies in emotion regulation.

In summary, we can divide the neural mechanism of love–envy into three levels: "love–envy elicitation," "evaluation of emotional information," and "cognitive control and emotion regulation" (Figure 6).

In terms of the mechanisms that generate the three types of envy, each of them is unique when it comes to the perception of stimuli in a context. As we can see from the neural mechanism models of social comparison envy and love envy, compared to the trait (personality) envy, social comparison envy and love–envy as two types of state emotions generated by specific experimental tasks, although using different types of experimental tasks, i.e., inducing social comparison envy using the story context method and the money gain/loss game, and inducing love–envy using the infidelity contextual utterance task and the contextual imagery tasks, but they all fit their respective definitions. Among them, social comparison envy was associated with brain areas of self-reference and social cognition, and love–envy was associated with brain areas of sex, reward, and motivation.

In terms of the emotion regulation mechanisms of envy, the three types of envy share very similar neural mechanisms. The cognitive control model of emotion regulation suggests that emotion regulation arises during the process of emotion onset and that different emotion regulation occurs at different stages of emotion onset (Gross and James, 1998; Gross and Thompson, 2007). Among them, cognitive changes are formed before the formation of emotional response tendencies, which are prior-focused emotion regulation and exhibit cognitive reappraisal emotion regulation strategies; response adjustments are made after the formation of emotional response tendencies, which are response-focused emotion regulation and exhibit expression inhibition of emotion regulation strategies. A single-unit analysis found that all three envy types induced brain regions associated with cognitive reappraisal and expression-inhibiting emotion regulation strategies, including the middle frontal gyrus, the inferior frontal gyrus, and the slope of the cerebellum. It is evident that when people develop envy, they use cognitive reappraisal to understand the adverse emotional event more positively or to rationalize the emotional event. Expressive inhibition is also used to mobilize self-control and to initiate self-control processes to inhibit one's emotional behavior.

4.2 Joint analysis: the relationship of neural mechanisms between different types of envy

Joint analysis showed that the fusion of trait (personality) and social comparison envy yielded eight clusters. The key brain regions mainly activated included the frontal sub-gyrus, inferior parietal lobule, inferior frontal gyrus, precuneus, paracentral lobule, declive, posterior lobule, and extra-nuclear lobule. The single-unit analysis shows that both envy types have sub-parietal lobules closely related to cognitive control and emotion regulation. Simultaneously, the inferior frontal gyrus influences mental reappraisal and expression inhibition in emotion regulation strategies. The expression suppression strategy triggers the involvement of the declive in the suppression of facial expressions associated with trait (personality) envy and social comparison envy. It was found that the precuneus, activated significantly after a joint analysis of the two envy types, was not activated considerably during a single meta-analysis of trait (personality) envy but was activated substantially during a single meta-analysis of social comparison envy. The precuneus is also involved in self-information processing related to the self (Northoff et al., 2006). The functional magnetic resonance imaging study of trait (personality) envy included in this study was measured by a scale with self-relevant items, such as "I feel jealous every day, and the feeling of envy torments me constantly" from the Dispositional Envy Scale. Noteworthy, trait (personality) envy may activate the precuneus lobe.

However, in contrast to social comparison envy, which is based on the contextual characteristics of upward social comparison triggering envy, the precuneus activation of the trait (personality) envy originates only from the measurement modality and not from the type of envy itself. Therefore, in the single meta-analysis, precuneus activation was insignificant within the activated brain regions for trait (personality) envy compared with other brain regions. The joint analysis also identified significantly activated brain regions not found in either envy type in the single meta-analysis: frontal sub-gyrus, paracentral lobule, posterior lobe, and extra-nuclear lobule. Among them, the brain regions associated with emotion regulation and cognition are the frontal sub-Gyral and posterior lobes (lobules VI and VII; Phillips et al., 2008; Mu et al., 2022), the paracentral lobule is involved in self-related information processing (Yang, 2016), and the lateral lobule nucleus cluster includes the lateral amygdala, basal amygdala, and parabasal amygdala, which are considered the main structures that provide information for emotion perception (Han, 2016). They were all associated with trait (personality) envy and social comparison envy; however, their role as a single envy type was insignificant.

On the one hand, it may be that other brain regions with the same function (e.g., sub-parietal lobule, precuneus, and lentiform nucleus) are relatively overshadowed by the more significant effect sizes. On the other hand, the paracentral envy function of processing information related to the self is not substantial in the trait (personality) envy, as it only originates from the measurement modality and not from the envy type itself. The function of the amygdala in providing emotion perception information in social comparison envy was more often completed when brain regions associated with social comparison were involved in the evoked paradigm and thus was not significant.

The joint analysis showed that trait (personality) envy and love–envy fusion yielded eight clusters. Critical brain regions that were mainly activated included the extra-nuclear lobule, lentiform nucleus, paracentral lobule, cingulate gyrus, inferior frontal gyrus, supramarginal gyrus, and middle frontal gyrus. The single meta-analysis clearly showed that both envy types have cingulate gyri that carry out emotional information responses and encode information about emotionally salient events, which influences the cognitive reappraisal strategy, and the inferior frontal gyrus, which expresses the inhibition strategy in the emotion regulation strategy. It was found that the lentiform nucleus and middle frontal gyrus, activated significantly when both envy types were analyzed jointly, were not markedly activated in the love–envy single meta-analysis but were activated mainly in the trait (personality) envy single meta-analysis. The role of the lentiform nucleus in the perception of negative emotions in love–envy was more often completed when brain regions associated with romantic love were involved in the evoked paradigm and, therefore, was not significant. The middle frontal gyrus has the same function as the inferior frontal gyrus. The inferior frontal gyrus's effect on love–envy may be more meaningful and relatively overshadow the impact of the middle frontal gyrus; therefore, it is not essential. The joint analysis also identified significantly activated brain regions not found in either envy type during the single meta-analysis, namely the paracentral lobule, extra-nuclear lobule, and limbic supramarginal gyrus. The paracentral lobule is involved in information processing related to the self; the amygdala in the lateral lobule cluster provides information on emotion perception (Han, 2016), and the supramarginal gyrus, a component of the inferior parietal lobule, is closely related to cognitive control and emotion regulation in common with it (Caspers et al., 2006). They are associated with both trait (personality) envy and love–envy; however, their role as a single envy type is insignificant.

On the one hand, other brain regions with the same function (e.g., the inferior parietal lobule) may be relatively masked by the more significant effect. On the other hand, the paracentral lobule's function of processing information related to the self in the trait (personality) envy originates only from how it is measured and not from the envy type itself and is therefore not significant. The function of the amygdala in providing information about emotional perception is often accomplished in love–envy when the brain regions associated with romantic love are involved in the evoked paradigm and are therefore not significant.

The joint analysis results showed that 10 clusters were obtained for social comparison envy and love–envy fusion. The key centrally activated brain regions included the anterior cingulate gyrus, insula, supramarginal gyrus, inferior frontal gyrus, cingulate gyrus, declive, middle frontal gyrus, temporal lobe sub-Gyral, and extra-nuclear lobule. The single meta-analysis results clearly showed that both envy types influenced the cognitive reappraisal strategy of emotion regulation and expression inhibition strategy of the inferior frontal gyrus. It was found that the anterior cingulate gyrus, middle frontal gyrus, and declive, significantly activated after the joint analysis of both envy types, were not particularly activated during the single meta-analysis of love–envy. Nonetheless, they were activated considerably during the single meta-analysis of social comparison envy. The anterior cingulate gyrus is closely associated with cognitive control (Meldrum et al., 2018). In love–envy, the insula, superior frontal gyrus, medial frontal gyrus, and inferior parietal lobule have similar functions. The effect sizes of these brain regions may be more significant and mask the role of the anterior cingulate gyrus; therefore, they are not necessary. The middle frontal gyrus has the same function as the inferior frontal gyrus. However, the inferior frontal gyrus's effect is possibly more significant in love–envy and overshadows the impact of the middle frontal gyrus; thus, it is not substantial.

The declive is associated with the "inhibitory" emotion regulation strategy. It has been found that expressing love–envy is more acceptable than expressing social comparison envy (Zhang et al., 2011). Individuals are less likely to use the "inhibitory" strategy for emotion regulation after love–envy is induced. The possibility of using the "expression inhibition" strategy for emotion regulation after the induction of love–envy was low and, therefore, insignificant. Simultaneously, the cingulate gyrus and insula, activated significantly after joint analysis of the two envy types, did not start considerably during the social comparison envy single meta-analysis but started especially during the love–envy single meta-analysis. The cingulate gyrus, associated with emotional information responses and information encoding emotionally salient events, was similarly activated during the social comparison envy evocation. However, in the single meta-analysis, it was found that the precentral gyrus inhibited the cingulate gyrus from identifying negative emotions more significantly in social comparison envy, which may have caused the cingulate gyrus to be insignificant in the single meta-analysis of social comparison envy and is consistent with the preference of social comparison envy for "inhibitory" emotion regulation strategies. The joint analysis also identified significantly activated brain regions that were not found in either type of envy in a single meta-analysis, namely the extra-nuclear lobule, supramarginal gyrus, and temporal lobe

sub-gyrus, where the amygdala in the lateral lobule nucleus provides emotional perception information (Han, 2016) and the supramarginal gyrus is closely related to cognitive control and emotion regulation (Caspers et al., 2006) and the temporal lobe is involved in cognitive information processing, situational memory encoding, and extraction processes (Chong et al., 2020). On the one hand, it may be that other brain regions with similar functions (e.g., inferior parietal lobule and superior temporal gyrus) were obscured by a more significant effect size. On the other hand, possibly, the role of the amygdala in providing emotional perception information was already completed when brain regions associated with social comparison in social comparison envy and romantic love–envy were activated in the evoked paradigm, which may explain why the amygdala was not found to be not significant in these contexts.

4.3 Contrasting analysis: the relationship of neural mechanisms between different types of envy

The results of the comparative analysis showed no activation clusters in the comparisons of the three types of envy. Possible reasons for this are as follows: first, the inclusion criteria for the ALE study literature are somewhat subjective. Second, according to the inclusion criteria, fewer papers met the inclusion criteria in this study, and the statistical validity may be weak, resulting in insignificant differences among the three envy types.

4.4 Limitations and outlook

There are some limitations to our work. First, limited by the number of existing studies, we did not find significant differences between the three types of envy. With the abundance of related studies, it is possible to clarify the characteristics of the three envy types using only the corresponding neural processing mechanisms in the future. Second, there is the problem of publication bias that should be addressed. Coordinate- based neuroimaging meta-analyses test for spatial convergence of effects across experiments with the null-hypothesis of random spatial convergence (Rottschy et al., 2012). Thus a limitation of most coordinate-based algorithms is that they are insensitive to non-significant results and publication bias may go unnoticed. Most of the articles related to ALE meta-analysis have not been tested for publication bias. This may be due to the inability to perform traditional publication bias tests using coordinates. In the study, we uses the main effect size of each literature to judge the problem of publication bias. There may be limitations to this approach. Third, unfortunately, there is currently no option for correlation analysis in GingerALE. Also, the small number of included literature is the impossibility to not only calculate one main meta-analysis, but rather also sub-analyses which may focus on more specialized processes (e.g., different paradigm classes) or groups (e.g., different samples). Due to this reason, we cannot control variables sufficiently to minimize the influence of potential factors on neuropsychological mechanisms. Finally, in addition to the three common types of envy in this study, researchers have focused on other types of envy, such as good-intentioned envy, as proposed by the dual structural theory of envy (Crusius et al., 2020). Regarding motivational and behavioral tendencies, when confronted with the envied person's superiority, individuals with good-intent envy will generate positive motivation that drives them to improve themselves through efforts; experienced envy or attributed envy based on emotional self-bias theory, among others. However, we did not include these newer types of envy in our literature inclusion because there has not been sufficient correlational research on fMRI to support the meta-analysis.

In the future, there is much work should to be done. First, our meta-analysis focused on the neural mechanisms of envy in healthy participants only. However, research has also been conducted on the neural mechanisms of envy in populations with autism, juvenile delinquents, and others (Daniel et al., 2021; Sol et al., 2023). Therefore, an important direction for future research is the translation of basic research on the generation and regulation of envy to understanding the full range of normal to abnormal differences in emotional generation and regulatory ability of envy. This is critical both for understanding the mechanisms underlying this variability and for testing the boundaries of basic models of envy generative and regulatory mechanisms. Second, one domain in which this will prove important is understanding how our envy changes as we grow from childhood through adolescence into adulthood and old age. The age of the participants in the literature included in this paper ranged from 17 to 27 years old. On the one hand, there is growing evidence that childhood and adolescence are critical times for the development of the envy regulatory abilities needed to adaptively regulate affective impulses and the deleterious offensive behavior they can promote (McRae et al., 2012). One the other hand, while physical health and cognitive abilities tend to decline with age (Grady, 2008), older adults report more emotional stability and a greater ratio of positive to negative experiences in their daily life, with the extent of positive emotion predicting longevity (Carstensen and Mikels, 2005). One conundrum to resolve here will be the apparent dependence of emotion regulation on the same kinds of prefrontal control systems that decline with age. This raises the question of how regulatory abilities improve as the underlying neural machinery declines (Ochsner et al., 2012). Early results suggest that it may depend on the strategies older adults deploy, with spared or greater regulatory ability shown for strategies and tactics that fit with long-term goals and have become habitual (Ochsner and Gross, 2008). Third, an important goal for future research will be to understand how potential dysfunction in the mechanisms of envy generation and regulation may underlie various forms of psychiatric and substance use disorders. This future direction is being pursued in studies across various disorders, ranging from delusional symptoms to depression and autism. These studies can be useful in two ways. First they may show disorder-specific patterns of altered function in control and affect systems. Second, imaging methods for studying emotion regulation may be used before and after treatment regimes as predictors of and markers of improvement.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

SD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. QL: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. HC: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. WZ: Conceptualization, Funding acquisition, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Social Science Fund of China (21CSH088): "Research on key elements and path optimization of improving the intervention mechanisms of social psychological crises under a new development pattern" and Zhejiang University of Technology Basic Research Operations Fund - Outstanding Young Scholars Special (GB202202005). This work was also supported by the Ministry of Education in China (MOE) Project of Humanities and Social Sciences (project no. 19YJC190028) and the Fundamental Research Funds for the Central Universities, Xidian University (ZYTS23130). The authors would like to express their gratitude for them.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anderson, C., Hildreth, J. A. D., and Howland, L. (2015). Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychol. Bull.* 141, 574–601. doi: 10.1037/a0038781

Arsalidou, M., Vijayarajah, S., and Sharaev, M. (2020). Basal ganglia lateralization in different types of reward. *Brain Imaging Behav.* 14, 2618–2646. doi: 10.1007/s11682-019-00215-3

Bartz, J. A., Zaki, J., Bolger, N., and Ochsner, K. N. (2011). Social effects of oxytocin in humans: context and person matter. *Trends Cogn. Sci.* 15, 301–309. doi: 10.1016/j.tics.2011.05.002

Bauminger, N. (2004). The expression and understanding of jealousy in children with autism. *Dev. Psychopathol.* 16, 157–177. doi: 10.1017/s0954579404044451

Belk, R. W. (1985). Materialism: trait aspects of living in the material world. *J. Consum. Res.* 12, 265–280. doi: 10.1086/208515

Brennan, M., Kerstin, B., Dar, M., and Christian, V. S. (2020). Individual differences in envy experienced through perspective–taking involves functional connectivity of the superior frontal gyrus. *Cogn. Affect. Behav. Neurosci.* 20, 783–797. doi: 10.3758/s13415-020-00802-8

Bringle, R. G. (1981). Conceptualizing jealousy as a disposition. *J. Fam. Econ. Iss.* 4, 274–290. doi: 10.1007/BF01257941

Bringle, R. G. (1991). *Psychosocial Aspects of Jealousy: A Transactional Model. In Salovey P. The Psychology of Jealousy and Envy*. The Guilford Press, 103–131.

Bringle, R. G., Roach, S., Andier, C., and Evenbeck, S. (1979). Measuring the intensity of jealous reactions. *Catalog Select Doc. Psychol.* 9, 23–24.

Buckner, R. L., and Carroll, D. C. (2007). Self-projection and the brain. *Trends Cogn. Sci.* 11, 49–57. doi: 10.1016/j.tics.2006.11.004

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., et al. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cereb. Cortex* 24, 2981–2990. doi: 10.1093/cercor/bht154

Bush, G., Luu, P., and Posner, M. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends Cogn. Sci.* 4, 215–222. doi: 10.1016/S1364-6613(00)01483-2

Button, K. S., Ioannidis, J. P. A., and Mokrysz, C. (2013). Erratum: power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475

Camicioli, R. (2011). Othello syndrome – at the interface of neurology and psychiatry. *Nat. Rev. Neurol.* 7, 477–478. doi: 10.1038/nrneurol.2011.123

Carstensen, L. L., and Mikels, J. A. (2005). At the intersection of emotion and cognition: aging and the positivity effect. *Curr. Dir. Psychol. Sci.* 14, 117–121. doi: 10.1111/j.0963-7214.2005.00348.x

Caspers, S., Geyer, S., Schleicher, A., and Zilles, K. (2006). The human inferior parietal cortex: Cytoarchitectonic parcellation and interindividual variability. *NeuroImage* 33, 430–448. doi: 10.1016/j.neuroimage.2006.06.054

Cauda, F., Costa, T., Diana, M. E. T., and Vercelli, A. (2012). Meta-analytic clustering of the insular cortex. *Neuro Image* 62, 343–355. doi: 10.1016/j.neuroimage.2012.04.012

Cavanna, A. E., and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* 129, 564–583. doi: 10.1093/brain/awl004

Chen, M. (2017). *The Effect of Oxytocin on Decision-Making Behavior and Its Neural Basis (Master's Thesis)*. Chongqing, China: Southwest University.

Chong, T. W. H., Curran, E., Ellis, K. A., and Nicola, T. L. (2020). Physical activity for older Australians with mild cognitive impairment or subjective cognitive decline – a narrative review to support guideline development. *J. Sci. Med. Sport* 23, 913–920. doi: 10.1016/j.jsams.2020.03.003

Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness). *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Crusius, J., Gonzalez, M. F., Lange, J., and Cohen-Charash, Y. (2020). Envy: an adversarial review and comparison of two competing views. *Emot. Rev.* 12, 3–21. doi: 10.1177/1754073919873131

Crusius, J., and Lange, J. (2016). *How Do People Respond to Threatened Social Status? Moderators of Benign Versus Malicious*. New York: Oxford University Press.

Dandurand, C., and Lafontaine, M. (2014). Jealousy and couple satisfaction: a romantic attachment perspective. *Marriage Fam. Rev.* 50, 154–173. doi: 10.1080/01494929.2013.879549

Daniel, F., Agustín, I., Hernando, S. G., Michel, P. S., Claudia, I., Mariana, P., et al. (2021). Neuroanatomy of complex social emotion dysregulation in adolescent offenders. *Cogn. Affect. Behav. Neurosci.* 21, 1083–1100. doi: 10.3758/s13415-021-00903-y

Deans, H., and Bhogal, M. S. (2019). Perpetrating cyber dating abuse: a brief report on the role of aggression, romantic jealousy and gender. *Curr. Psychol.* 38, 1077–1082. doi: 10.1007/s12144-017-9715-4

Deng, D., Liao, H., Duan, G., He, Q., Pang, Y., Liu, H., et al. (2015). Changes in low–frequency amplitude in functional brain areas after acupuncture at Baihui point in patients with major depressive disorder. *J. Clin. Radiol.* 34, 884–888. doi: 10.13437/j.cnki.jcr.2015.06.007

Dong, P., Cui, F., Tan, Z., Jiang, G., Zhang, H., and Zou, Y. (2010). Functional magnetic resonance imaging study of the effect of acupuncture on the remodeling of brain function in patients with spasticity after cerebral infarction. *Chin. J. Rehabil. Med.* 25, 507–513. doi: 10.3969/j.issn.1001-1242.2010.06.004

Dvash, J., Gilam, G., Ben-Ze'ev, A., Hendler, T., and Shamay-Tsoory, S. G. (2010). The envious brain: the neural basis of social comparison. *Hum. Brain Mapp.* 31, 1741–1750. doi: 10.1002/hbm.20972

Eickhoff, S. B., Laird, A. R., Fox, P. M., Lancaster, J. L., and Fox, P. T. (2016). Implementation errors in the ginger ALE software: description and recommendations. *Hum. Brain Mapp.* 38, 7–11. doi: 10.1002/hbm.23342

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* 113, 7900–7905. doi: 10.1073/pnas.1602413113

Ellenbogen, M. A., Linnen, A. M., Grumet, R., Cardoso, C., and Joober, R. (2012). The acute effects of intranasal oxytocin on automatic and effortful attentional shifting to emotional faces. *Psychophysiology* 49, 128–137. doi: 10.1111/j.1469-8986.2011.01278.x

Enatsu, R., Bulacio, J., Nair, D. R., Bingaman, W., Najm, I., and Gonzalez-Martinez, J. (2014). Posterior cingulate epilepsy: clinical and neurophysiological analysis. *J. Neurol. Neurosurg. Psychiatry* 85, 44–50. doi: 10.1136/jnnp-2013-305604

Festinger, L. (1954). A theory of social comparison processes. *Hum. Relat.* 7, 117–140. doi: 10.1177/001872675400700202

Fiske, S. T. (2010). Envy up, scorn down: how comparison divides us. *Am. Psychol.* 65, 698–706. doi: 10.1037/0003-066X.65.8.698

Garnefski, N., and Kraaij, V. (2006). Cognitive emotion regulation questionnaire-development of a short 18-item version (CERQ-short). *Curr. Dir. Psychol. Sci.* 17, 153–158.

Gibbons, F. X., and Buunk, B. P. (1999). Individual differences in social comparison: development of a scale of social comparison orientation. *J. Pers. Soc. Psychol.* 76, 129–142. doi: 10.1037/0022-3514.76.1.129

Gold, B. T. (1996). Enviousness and its relationship to maladjustment and psychopathology. *Personal. Individ. Differ.* 21, 311–321. doi: 10.1016/0191-8869(96)00081-5

Grady, C. L. (2008). Cognitive neuroscience of aging. *Ann. N. Y. Acad. Sci.* 1124, 124–144. doi: 10.1196/annals.1440.009

Gross, J. J., and James, J. (1998). The emerging field of emotion regulation: an integrative review. *Rev. Gen. Psychol.* 2, 271–299. doi: 10.1037/1089-2680.2.3.271

Gross, J. J., and Thompson, R. A. (2007). *Handbook of Emotion Regulation*. New York: Guiford Press.

Han, Y. (2016). *Effects of Oxytocin Action on Central Amygdala on Nociception in Rats (Master's thesis)*. Yunnan, China: Yunnan University.

Hupka, R. B., and Bachelor, B. (1979). Validation of a Scale to Measure Romantic Jealousy. Paper Presented at the Annual Convention of the Western Psychological Association, San Diego, CA, 8–10.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive science: detection, prevalence, and prevention. *Trends Cogn. Sci.* 18, 235–241. doi: 10.1016/j.tics.2014.02.010

Kar, H. L., and O'Leary, K. D. (2013). Patterns of psychological aggression, dominance, and jealousy within marriage. *J. Fam. Violence* 28, 109–119. doi: 10.1007/s10896-012-9492-7

Katrin, P., Dirk, S., Monika, E., and René, H. (2015). The influence of oxytocin on volitional and emotional ambivalence. *Soc. Cogn. Affect. Neurosci.* 10, 987–993. doi: 10.1093/scan/nsu147

Laird, A. R., Robinson, J. L., McMillan, K. M., Tordesillas-Gutierrez, D., Moran, S. T., Gonzales, S. M., et al. (2010). Comparison of the disparity between Talairach and MNI coordinates in functional neuroimaging data: validation of the Lancaster transform. *NeuroImage* 51, 677–683. doi: 10.1016/j.neuroimage.2010.02.048

Lancaster, J. L., Tordesillas-Gutierrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., et al. (2007). Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum. Brain Mapp.* 28, 1194–1205. doi: 10.1002/hbm.20345

Lange, J., and Crusius, J. (2015a). Dispositional envy revisited: unraveling the motivational dynamics of benign and malicious envy. *Personal. Soc. Psychol. Bull.* 41, 284–294. doi: 10.1177/0146167214564959

Lange, J., and Crusius, J. (2015b). The tango of two deadly sins: the social–functional relation of envy and pride. *J. Pers. Soc. Psychol.* 109, 453–472. doi: 10.1037/pspi0000026

Lim, E. C., Tan, J. J., Ong, B. K., and Smith, W. (2004). Generalized myoclonus evolving into epilepsia partialis continua due to a cingulated gyrus lesion: case report and review of the literature. *Parkinsonism Relat. Disord.* 10, 447–449. doi: 10.1016/j.parkreldis.2004.04.010

Liu, J. C., Ran, G., and Zhang, Q. (2022). Neural activity of different emotion carriers and their similarities and differences-an ALE meta-analysis of brain imaging studies. *Adv. Psychol. Sci.* 30, 536–558. doi: 10.3724/SP.J.1042.2022.00536

Mathes, E. W. (1992). *Jealousy: The Psychological Data*. New York: University Press of America.

McRae, K., Gross, J. J., Weber, J., Robertson, E. R., Sokol-Hessner, P., Ray, R. D., et al. (2012). The development of emotion regulation: an fMRI study of cognitive reappraisal in children, adolescents and young adults. *Soc. Cogn. Affect. Neurosci.* 7, 11–22. doi: 10.1093/scan/nsr093

Meldrum, R. C., Elisa, M. T., Lora, M. C., and Mary, M. H. (2018). Brain activity, low self-control, and delinquency: an fMRI study of at-risk adolescents. *J. Crim. Just.* 56, 107–117. doi: 10.1016/j.jcrimjus.2017.07.007

Moraweta, C., Bode, S., Baudewig, J., and Heekeren, H. R. (2017). Effective amygdala-prefrontal connectivity predicts individual differences in successful emotion regulation. *Soc. Cogn. Affect. Neurosci.* 12, 569–585. doi: 10.1093/scan/nsw169

Morawetz, C., Bode, S., Derntl, B., and Hauke, R. H. (2017). The effect of strategies, goals and stimulus material on the neural mechanisms of emotion regulation: a meta-analysis of fMRI studies. *Neurosci. Biobehav. Rev.* 72, 111–128. doi: 10.1016/j.neubiorev.2016.11.014

Mu, H., Jiang, R., Wang, Y., Ju, Y., and Zhao, X. (2022). Progress in the study of non–motor dysfunction after cerebellar stroke. *Chin. J. Stroke* 17, 905–910. doi: 10.3969/j.issn.1673-5765.2022.08.018

Mufson, E. J., and Mesulam, M. M. (1982). Insula of the old world monkey. II: afferent cortical input and comments on the claustrum. *J. Comp. Neurol.* 212, 23–37. doi: 10.1002/cne.902120103

Müller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T., Radua, J., Mataix-Cols, D., et al. (2018). Ten simple rules for neuroimaging meta-analysis. *Neurosci. Biobehav. Rev.* 84, 151–161. doi: 10.1016/j.neubiorev.2017.11.012

Nadine, S., Silvia, O. S., Gerald, E., Aylin, T., Jürgen, T., Katja, B., et al. (2019). The obsessions of the green–eyed monster: jealousy and the female brain. *Sex. Relatsh. Ther.* 36, 91–105. doi: 10.1080/14681994.2019.1615047

Neal, A. M., and Lemay, E. P. (2014). How partners' temptation leads to their heightened commitment: the interpersonal regulation of infidelity threats. *J. Soc. Pers. Relat.* 31, 938–957. doi: 10.1177/0265407513512745

Nimarko, A. F., Garrett, A. S., Carlson, G. A., and Singh, M. K. (2019). Neural correlates of emotion processing predict resilience in youth at familial risk for mood disorders. *Dev. Psychopathol.* 31, 1037–1052. doi: 10.1017/S0954579419000579

Northoff, G., Heinzel, A., Greck, M., Bermpohl, F., Dobrowolny, H., and Panksepp, J. (2006). Self-referential processing in our brain--a meta-analysis of imaging studies on the self. *NeuroImage* 31, 440–457. doi: 10.1016/j.neuroimage.2005.12.002

Ochsner, K. N., and Gross, J. J. (2007). *Handbook of Emotion Regualtion*. New York: Guiford Press.

Ochsner, K. N., and Gross, J. J. (2008). Cognitive emotion regulation: insights from social cognitive and affective neuroscience. *Curr. Direct. Psychol. Sci.* 17, 153–158. doi: 10.1111/j.1467-8721.2008.00566.x

Ochsner, K. N., Silvers, J. A., and Buhle, J. T. (2012). Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion. *Ann. N. Y. Acad. Sci.* 1251, 1–24. doi: 10.1111/j.1749-6632.2012.06751.x

Olff, M., Frijling, J. L., Kubzansky, L. D., Bradley, B., Ellenbogen, M. A., Cardoso, C., et al. (2013). The role of oxytocin in social bonding, stress regulation and mental health: an update on the moderating effects of context and interindividual differences. *Psychoneuroendocrinology* 38, 1883–1894. doi: 10.1016/j.psyneuen.2013.06.019

Pang, J. (2016). A localized interpretation of contemporary college students' jealousy. *J. Tianjin Textbook Inst.* 28, 57–60. doi: 10.16137/j.cnki.cn12-1303/g4.2016.04.015

Pfeiffer, S. M., and Wong, P. T. P. (1989). Multidimensional jealousy. *J. Soc. Pers. Relat.* 6, 181–196. doi: 10.1177/026540758900600203

Phillips, M. L., Ladouceur, C. D., and Drevets, W. C. (2008). A neural model of voluntary and automatic emotion regulation: implications for understanding the pathophysiology and neurodevelopment of bipolar disorder. *Mol. Psychiatry* 13, 833–857. doi: 10.1038/mp.2008.65

Protasi, S. (2016). Varieties of envy. *Philos. Psychol.* 29, 535–549. doi: 10.1080/09515089.2015.1115475

Redouté, J., Stoléru, S., Grégoire, M. C., Costes, N., Cinotti, L., Lavenne, F., et al. (2000). Brain processing of visual sexual stimuli in human males. *Hum. Brain Mapp.* 11, 162–177. doi: 10.1002/1097-0193(200011)11:3<162::AID-HBM30>3.0.CO;2-A

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., et al. (2012). Modelling neural correlates of working memory: a coordinate-based meta-analysis. *Neuro Image* 60, 830–846. doi: 10.1016/j.neuroimage.2011.11.050

Salovey, P., and Rodin, J. (1984). Some antecedents and consequences of social-comparison jealousy. *J. Pers. Soc. Psychol.* 47, 780–792. doi: 10.1037/0022-3514.47.4.780

Sheline, Y. I., Price, J. L., Yan, Z., and Mintun, M. A. (2010). Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc. Natl. Acad. Sci.* 107, 11020–11025. doi: 10.1073/pnas.1000446107

Silver, M., and Sabini, J. (1978). The social construction of envy. *J. Theory Soc. Behav.* 8, 313–331.

Smith, R. H., Parrott, W. G., Diener, E. F., Hoyle, R. H., and Kim, S. H. (1999). Dispositional envy. *Personal. Soc. Psychol. Bull.* 25, 1007–1020. doi: 10.1177/01461672992511008

Sol, F., Jorge, A., Joaquín, M., Matías, C., Agustín, I., and Sandra, B. (2023). Overactivation of posterior insular, postcentral, and temporal regions during the preserved experience of envy in autism. *Eur. J. Neurosci.* 57, 705–717. doi: 10.1111/ejn.15911

Song, L., Meng, J., Liu, Q., Huo, T., Zhu, X., Li, Y., et al. (2019). Polygenic score of subjective well-being is associated with the brain morphology in superior temporal Gyrus and insula. *Neuroscience* 414, 210–218. doi: 10.1016/j.neuroscience.2019.05.055

Song, H., Zou, Z., Kou, J., and Zhang, X. (2015). Love-related changes in the brain: a resting-state functional magnetic resonance imaging study. *Front. Hum. Neurosci.* 9:71. doi: 10.3389/fnhum.2015.00071

Steinbeis, N., and Singer, T. (2013). The effects of social comparison on social emotions and behavior during childhood: the ontogeny of envy and Schadenfreude predicts developmental changes in equity–related decisions. *J. Exp. Child Psychol.* 115, 198–209. doi: 10.1016/j.jecp.2012.11.009

Sun, Y., Yu, H., Chen, J., and Shi, J. (2016). Neural substrates and behavioral profiles of romantic jealousy and its temporal dynamics. *Sci. Rep.* 6:27469. doi: 10.1038/srep27469

Takahashi, H., Matsuura, M., Yahata, N., and Okubo, Y. (2006). Men and women show distinct brain activations during imagery of sexual and emotional infidelity. *NeuroImage* 32, 1299–1307. doi: 10.1016/j.neuroimage.2006.05.049

Tanaka, T., Nishimura, F., Kakiuchi, C., Kasai, K., Kimura, M., and Haruno, M. (2019). Interactive effects of OXTR and GAD1 on envy-associated behaviors and neural responses. *PLoS One* 14:e0210493. doi: 10.1371/journal.pone.0210493

Thomas, A. F., and Joseph, R. B. (2016). Negative mood regulation expectancies moderate the association between happiness emotion goals and depressive symptoms. *Personal. Individ. Differ.* 100, 23–27. doi: 10.1016/j.paid.2015.08.010

Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., and Fox, P. (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum. Brain Mapp.* 33, 1–13. doi: 10.1002/hbm.21186

Villablanca, J. R. (2010). Why do we have a caudate nucleus? *Acta Neurobiol. Exp.* 70, 95–105. doi: 10.55782/ane-2010-1778

Wager, T. D., Lindquist, M., and Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* 2, 150–158. doi: 10.1093/scan/nsm015

Xiang, Y., Kong, F., Wen, X., and Mo, L. (2016). Neural correlates of envy: regional homogeneity of resting–state brain activity predicts dispositional envy. *Neuro Image* 142, 225–230. doi: 10.1016/j.neuroimage.2016.08.003

Xiang, Y., Zhao, S., Wang, H., and Mo, L. (2017). Examining brain structures associated with dispositional envy and the mediation role of emotional intelligence. *Sci. Rep.* 7:39947.

Yang, L. (2016). *The Effect and Its Neural Mechanism of the Emotion and Self-Relevance on Episodic Future Thing (Master's thesis)*. Chongqing, China: Southwest University.

Yarkoni, T. (2009). Big correlations in little studies: inflated fMRI correlations reflect low statistical power – commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* 4, 294–298. doi: 10.1111/j.1745-6924.2009.01127.x

Yarkoni, T., Poldrack, R. A., Nichols, T. S., van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670. doi: 10.1038/nmeth.1635

Zeelenberg, M., and Pieters, R. (2007). A theory of regret regulation 1.0. *J. Consum. Psychol.* 17, 3–18. doi: 10.1207/s15327663jcp1701_3

Zhang, Y. J., Xu, Y. L., and Chen, F. C. (2011). Exploring the relationship between social comparison jealousy and love jealousy. *J. Sichuan Vocat. Tech. Coll.* 21:35–36+58. doi: 10.13974/j.cnki.51-1645/z.2011.05.017

Zheng, X. (2021). *Study the Neural Mechanism of Jealousy Trait and the Regulatory Role of Oxytocin (Doctoral Dissertation)*. Sichuang, China: University of Electronic Science and Technology.

Zheng, X., Luo, L., Li, J., and Kendrick, K. M. (2019). A dimensional approach to jealousy reveals enhanced frontostriatal, insula, and limbic responses to angry faces. *Brain Struct. Funct.* 224, 3201–3212. doi: 10.1007/s00429-019-01958-x

Zhu, C. (2002). *Neuroanatomy*. Beijing: People's Health Publishing House.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 21 March 2024
DOI 10.3389/fpsyg.2024.1362064

OPEN ACCESS

EDITED BY
Carmelo Mario Vicario,
University of Messina, Italy

REVIEWED BY
Peggy Mason,
The University of Chicago, United States
Adriana Foster,
Nova Southeastern University, United States

*CORRESPONDENCE
David Martínez-Pernía
✉ david.martinez@uai.cl

RECEIVED 05 January 2024
ACCEPTED 27 February 2024
PUBLISHED 21 March 2024

CITATION
Troncoso A, Blanco K, Rivera-Rei Á and Martínez-Pernía D (2024) Empathy bodyssence: temporal dynamics of sensorimotor and physiological responses and the subjective experience in synchrony with the other's suffering.
Front. Psychol. 15:1362064.
doi: 10.3389/fpsyg.2024.1362064

Empathy bodyssence: temporal dynamics of sensorimotor and physiological responses and the subjective experience in synchrony with the other's suffering

Alejandro Troncoso[1], Kevin Blanco[1], Álvaro Rivera-Rei[1] and David Martínez-Pernía[1,2]*

[1]Center for Social and Cognitive Neuroscience, School of Psychology, Adolfo Ibáñez University, Santiago, Chile, [2]Geroscience Center for Health and Brain Metabolism (GERO), Santiago, Chile

Background: Empathy is foundational in our intersubjective interactions, connecting with others across bodily, emotional, and cognitive dimensions. Previous evidence suggests that observing individuals in painful situations elicits whole bodily responses, unveiling the interdependence of the body and empathy. Although the role of the body has been extensively described, the temporal structure of bodily responses and its association with the comprehension of subjective experiences remain unclear.

Objective: Building upon the enactive approach, our study introduces and examines "bodyssence," a neologism formed from "body" and "essence." Our primary goal is to analyze the temporal dynamics, physiological, and phenomenological elements in synchrony with the experiences of sportspersons suffering physical accidents.

Methods: Using the empirical 5E approach, a refinement of Varela's neurophenomenological program, we integrated both objective third-person measurements (postural sway, electrodermal response, and heart rate) and first-person descriptions (phenomenological data). Thirty-five participants watched videos of sportspersons experiencing physical accidents during extreme sports practice, as well as neutral videos, while standing on a force platform and wearing electrodermal and heart electrodes. Subsequently, micro-phenomenological interviews were conducted.

Results: Bodyssence is composed of three distinct temporal dynamics. Forefeel marks the commencement phase, encapsulating the body's pre-reflective consciousness as participants anticipate impending physical accidents involving extreme sportspersons, manifested through minimal postural movement and high heart rate. Fullfeel, capturing the zenith of empathetic engagement, is defined by profound negative emotions, and significant bodily and kinesthetic sensations, with this stage notably featuring an increase in postural movement alongside a reduction in heart rate. In the Reliefeel phase, participants report a decrease in emotional intensity, feeling a sense of relief, as their postural control starts to reach a state of equilibrium, and heart rate remaining low. Throughout these phases, the level of electrodermal activity consistently remains high.

Conclusion: This study through an enactive approach elucidates the temporal attunement of bodily experience to the pain experienced by others. The integration of both first and third-person perspectives through an empirical 5E approach reveals the intricate nature of bodyssence, offering an innovative approach to understanding the dynamic nature of empathy.

KEYWORDS

bodyssence, empathy, enaction, neurophenomenology, empirical 5E approach, first-person view, phenomenology, mixed-method study

1 Introduction

Empathy is a critical component of our intersubjective life, playing a vital role in our ability to connect with others on a bodily, emotional, and cognitive level. It refers to our basic capacity to share, feel, or recognize another person's world (Eklund and Meranius, 2021). This ability to resonate with others involves not only psychological processes but also whole-body mechanisms that enable us to experience the emotions of another person firsthand (Gallese, 2014; de Waal and Preston, 2017; Riečanský and Lamm, 2019). Additionally, these mechanisms allow us to respond to another's emotions through action (Gallese, 2014; de Waal and Preston, 2017; Riečanský and Lamm, 2019). This cycle of perception and action through the body has been a prominent focus in the field over the past few decades. Studies examining the role of the body have investigated movement responses, autonomic reactions, and the activation of various areas of the motor control system in response to images or videos of another person in pain [for reviews, see Riečanský and Lamm, 2019 and Troncoso et al., 2023].

For instance, physiological and motor changes have been reported during the observation of others' emotions (Hein et al., 2011; Gea et al., 2014; Lelard et al., 2014; Mouras and Lelard, 2021). Another observation found that viewing faces in pain leads to increased amplitude of body sway in subjects and correlates with higher empathic subjective scores, suggesting that changes in postural control may be associated with approach and cooperative responses (Gea et al., 2014; Lebert et al., 2020). Concerning cardiac activity, individuals' emotional regulation of empathy through up-regulation and down-regulation while watching emotional videos were linked to increased and decreased subjective scores of situational empathy and distinct changes in heart rate variability compared to individuals in the control condition (No-regulation) while participants viewed emotional videos (Jauniaux et al., 2020). Complementing these observations, studies employing neuroimaging techniques have demonstrated that these bodily responses are reflected in brain activity. Indeed, numerous neuroimaging studies have indicated that activity in the somatosensory and motor cortices, typically associated with first-hand pain, is also triggered in response to the pain of others (Bufalari et al., 2007; Lamm et al., 2007; Riečanský and Lamm, 2019). These activations occur at multiple levels of the nervous system, from the cerebral cortex down to the spinal cord (Riečanský and Lamm, 2019). These findings underscore the intricate interplay between the brain and body in empathic experiences, offering a tangible illustration of how the body is fundamental in the empathic experience.

While the study of the role of the body in empathy research has furnished valuable insights and heavily influenced empirical empathy research, there is an escalating need for a more integrated and holistic approach. Current methodologies, largely centered on correlating neurophysiological and subjective data (self-report questionnaires), may not fully capture the intrinsic dynamics of empathy within lived and situational contexts (Varela et al., 1991; Gallagher, 2003; De Jaegher and Di Paolo, 2007; Olivares et al., 2015). This is where the enactive approach, prioritizing the phenomenological experience and the mutual co-constitution of agent and environment, offers a unique and complementary lens for understanding empathy (Varela et al., 1991; Thompson, 2007; Colombetti, 2014; Newen et al., 2018; Troncoso et al., 2023). This approach proposes that the primary source of experience and understanding of others is through and with the body (Gallagher, 2012; Tanaka, 2015; Fuchs, 2017). Through our lived body we experience the physical expressions of others as meaningful actions that communicate their intentions, needs, and objectives within a shared context (Gallagher, 2012; Tanaka, 2015; Fuchs, 2017). In this sense, phenomenological descriptions have demonstrated the multidimensional and complex dynamic nature of subjective experiences of empathy and how they are formed by neurophysiological responses (Grice-Jackson et al., 2017a; Martínez-Pernía et al., 2023). For example, Grice-Jackson et al. (2017b) used bodily experience descriptions to distinguish different responders in a classic empathy for pain study. This phenomenological clustering revealed different functional brain connectivity between different kinds of conscious bodily feeling (Grice-Jackson et al., 2017b).

Building upon this foundational understanding, it is crucial to address the temporal dynamics of empathy. The enactive perspective posits that empathy is not just an individual experience but a collaborative process involving mutual interaction in a shared environment (Colombetti, 2014; Laroche et al., 2014; Fuchs, 2017). Empathy, therefore, is not a static response but a temporal dance of mutual adaptation where our reactions harmonize with the actions, emotions, and postures of those around us (Fuchs, 2013, 2017). Such dynamic interactions influence both parties involved as highlighted by studies showing spontaneous physiological coordination between individuals simply due to their co-presence (Golland et al., 2015). Emphasizing these temporal dynamics is paramount not only for a deeper understanding of the neurological facets of empathy but also for developing empathy theories (Golland et al., 2015).

In summary, the enactive approach offers a fresh and integrative perspective on empathy, highlighting the intricate interplay of embodiment, phenomenological insights, and their timely resonance with the environment. Despite the advancements in the enactive approach, the ever-evolving discipline of empathy research requires the ongoing development and refinement of concepts and more empirical evidence to support its theoretical formulations. Contemporary literature is replete with concepts such as the "lived body," "affordance,"

and "bodily resonance" seeking a more intricate and holistic understanding of the human being. Yet, there remains a significant empirical gap in understanding how these biological, subjective, and temporal processes converge. Seeking to bridge this chasm, we introduce the notion of "bodyssence." This neologism composed of the words "body" and "essence," represents a holistic exploration that bridges our neurophysiological responses with our subjective experiences. It underscores the dynamic interplay of these dimensions, both in the moment's immediacy and over time. This concept seeks to understand how our embodied experiences dynamically evolve and resonate with both our internal and external states in synchrony with the environment.

Therefore, in this research, our objective is to explore "bodyssence" more deeply within the context of empathy and to examine how the temporal evolution of motor, physiological, and phenomenological correlates, in synchrony with other, shapes our empathetic consciousness. Thus, we employed a refinement of Varela's neurophenomenological program (Varela, 1996) termed the empirical 5E approach (Troncoso et al., 2023), in which data from the sensorimotor system and physiological responses are collected using a mobile brain–body imaging system (MoBi) alongside the subjective experience. Participants were exposed to videos of people in physical accident situations (empathy for pain condition) and neutral videos that allowed the assessment of a baseline response from the participants (baseline condition). In the exposure, postural and physiological responses (electrocardiogram and electrodermal activity) were examined. Next, a micro-phenomenological interview was performed to explore the multi-layered dimensions (bodily sensations, emotions, and motivations) and temporal aspects of empathic experience. Ultimately, the integration of motor and physiological data will capture a multifaceted interplay of temporal, corporeal, and dynamic shifts as they relate to the subject of empathy, culminating in a comprehensive view of empathy for pain through our 'bodyssence' conceptualization. We hypothesize that a subjective climax experienced during the fall of sportspersons may be associated with an increase in anteroposterior movement, electrodermal activity, and heart rate. Additionally, following the peak climax, we hypothesize that all variables will decrease compared to the highest point, while the subjective experience will exhibit a decrease of anguish.

With the introduction of the term "bodyssence," we seek to clarify this intertwined terrain of empathy. "Bodyssence" not only offers a more detailed and comprehensive perspective on the interweaving of bodily, subjective processes in interaction with the environment but also establishes a clear research methodology that transitions from an enactive theoretical framework to a precise scientific mixed-method study (empirical 5E approach). With this initiative, we aim to energize the scientific community to embark on empirical research within the enactive perspective, especially since the majority of this field has remained theoretical. The scientific validity of this paradigm demands, and indeed compels both philosophers and neuroscientists alike, and associated disciplines, to uncover scientific findings that will make this vision enduring.

2 Methods

2.1 Participants

From September 2017 to January 2018, thirty-five adults participated in the study (19 female; mean age 30.18 ± 6.62 years; range 21–47 years, mean years of education = 17.17 ± 2.35)[1], all of whom were Latin American and Spanish-speaking. They were all healthy and did not have any cognitive or physical conditions that could affect their normal psychological and motor faculties. To corroborate that participants met the inclusion criteria, brief interview questionnaires were administered using the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005),the Beck's Depression Inventory (BDI-II, Beck et al., 1996), and State Trait Anxiety Inventory (STAI; Spielberger et al., 1970). The results of the questionnaires were as follows: MoCA = 28.57 ± 1.46; BDI-II = 5.4 ± 5.87; STAI = 49.4 ± 12.04.

All the participants signed informed consent. The study procedure was according to the Declaration of Helsinki principles. It was approved by the *"Scientific Ethics Committee of the Servicio de Salud Metropolitano Oriente"* and the *"Research in Humans Being Ethics Committee of the Medicine Faculty, Universidad de Chile."*

2.2 Construction and validation of the emotional stimuli

For the construction of the empathy for pain and baseline video conditions, 12 scenes were produced for each condition using audiovisual material found under Creative Commons licenses on the web. Each scene had an average duration of 7–11 s. The scenes for the pain condition included images of sportspersons suffering intense physical accidents during the practice of extreme sports (e.g., parkour, high mountain slackline, acrobatic snowboarding). Scenes of dismemberment, disfigurement, or death were not used. Stimuli with significant camera movements or vibrations were also excluded, as were scenes that produced saccadic eye movements. All scenes for the empathy for pain condition had a similar sequence: a sportsperson skillfully practices a sports activity (pre-fall); next, the sportsperson starts losing balance until they have a strong impact with the ground (fall); and finally, the sportsperson is seen moving after the impact (post-fall). In contrast, the baseline condition, consisting of images of domestic spaces, was established to gauge a fundamental neurophysiological response. This provides a foundational reference against which the heightened responses from the empathy for pain condition can be effectively compared over time.

After the empathy for pain and baseline were constructed, they were validated with 65 university students (38 female and 27 males; average age = 19.34 ± 1.56) using the Self-Assessment Manikin scale (Bradley and Lang, 1994), which evaluates valence, arousal, and dominance on a 9-point Likert scale. Higher scores indicate pleasant valence, greater arousal, and dominance of the situation; lower scores indicate unpleasant valence, less arousal, and dominance of control over the situation. Finally, two 60-s videos were constructed (empathy for pain condition and baseline condition) each containing seven scenes. Paired t-tests showed that the videos selected for the empathy for pain stimuli were assigned significantly lower valence (pain: mean = 3.77 ± 1.94; baseline: mean = 4.97 ± 2.39; t (64) = −7.24, p < 0.001), higher arousal (pain: mean = 6.40 ± 1.78; baseline:

1 Twenty-eight participants were extracted from our previous study (Martinez-Pernía et al., 2023). In our previous study we focused on exploring the structure of experiences (self-centered and other-centered empathy) based on non-temporal phenomenological analysis.

mean = 3.02 ± 2.24; t (64) = 24.91, $p < 0.001$), and lower dominance (pain: mean = 5.31 ± 2.68; baseline: mean = 7.66 ± 2.25; t (64) = −14.59, $p < 0.001$) than baseline stimuli. The empathy for pain video was previously constructed and validated by our group, and has been utilized in prior works (Martínez-Pernía et al., 2020, 2023; Pizarro et al., 2023).

2.3 Procedure

Participants were asked to maintain a quiet stance while standing on a force plate with hip-width feet positioned, arms rested alongside their body, and at a 1-meter distance from a 40-inch screen TV. Each condition video (empathy for pain and baseline) was randomly reproduced on the screen while the postural control data and physiological data were collected. The postural control data were collected by a Bertec FP4060-05-PT brand stabilometric platform (Bertec Corporation, Columbus, Ohio, USA). A BIOPAC MP150 data acquisition and analysis system with AcqKnowledge software (BIOPAC Systems, Inc.) was used for the integration of all stabilometry signals The electrocardiogram (ECG) was recorded using disposable snap ECG electrodes in a modified lead II configuration, with one electrode positioned beneath the right clavicle and the other near the lower ribs on the left side. Concurrently, the electrodermal activity (EDA) was measured using disposable snap electrodes attached to the palmar surface of the distal phalanges of the first and second fingers.

A script made in MATLAB (MathWorks, Natick, MA, United States) was used to present the stimulus on a screen (40 inches) and send signals to the AcqKnowledge software for synchronization of the stimulus presentation and the stabilometry and physiological data. The ECG and the EDA were also collected and synchronized with the stimulus presentation through the BIOPAC. Immediately after participants finished watching each video, the 9-point Likert scale of the Self-Assessment Manikin was administered to explore intensity and arousal. After, a researcher conducted a phenomenological interview focused on the pain condition.

2.4 The phenomenological interview

All interviews were conducted in Spanish by the same researcher, recorded using an audio device, and subsequently transcribed verbatim. At the beginning of each interview, participants were asked to describe the videos they found unpleasant and then select the one that represented the most intense overall experience. Next, we asked about the overall temporality of all videos to corroborate a similar pattern among the unpleasant videos. This approach was employed to facilitate a clearer recollection of the experience. Interviews were conducted following the micro-phenomenological interview (Petitmengin, 2006).

The interviews were focused on multidimensional experiences (bodily, affective, sensory, attentional, etc.) at specific moments, as well as on the fluctuations of these dimensions throughout the experience. A vital aspect of the interviews was maintaining a continuous awareness (by both the researcher and the participant) of the suspension of the "natural attitude," and judgmental stance toward one's own experience (Hamilton et al., 2019).

Initially, the interview commenced with a description of the interview objectives and the embodied approach to questions and answers, concentrating on the experiences related to the video. Next, the participant was prompted to evoke the video experience. This evocation principle was crucial to elicit the participants' pre-reflective descriptions and to vividly explore their past experiences (Petitmengin et al., 2018). The interview was developed by focusing on the synchronic aspect of the experience (e.g., "How do you feel about the video?" or "What is the feeling of tension like?") and the diachronic dimensions (e.g., "After the feeling of tension, what happens?" or "At what point do you feel the tension?"). Another characteristic of the interview procedure was to recapitulate the participant's responses to facilitate their recall.

2.5 Data analysis

2.5.1 Behavioral and physiological data analysis

The stabilometric force (Fx, Fy, Fz) and moment (Mx, My, Mz) components were collected at a sample rate of 125 Hz. The center-of-pressure (COP) was computed for the anteroposterior and mid-lateral directions. The COP series were filtered with an 8 Hz fourth-order lowpass Butterworth filter. The EDA data was sampled at 500 Hz. The phasic component was estimated using the convex optimization approach (Greco et al., 2016). The statistical analysis was performed on the mean phasic component of the EDA (mS). The ECG data was sampled at 500 Hz. The R peaks were identified in MATLAB with the peak finder function. Each variable was analyzed through seven temporal windows of 1 s. The selection of these windows began by identifying the moment of the athlete's fall. This event was identified as time 0. Next, three windows were taken before the athlete's fall and another four windows after the fall for 1 s each. The seven windows of the baseline condition were selected by taking the central temporal part of each scene. Subsequently, and in each window, the anteroposterior (AP) amplitude of the CoP was calculated. The AP amplitude was chosen based on the sensitivity demonstrated in previous studies that investigated postural responses in emotional research (Lelard et al., 2019). For the EDA and ECG, the phasic component and heart rhythm were analyzed, respectively.

2.5.2 First-person data analysis

To conduct the phenomenological analysis of the data, we used the descriptive phenomenological psychological method, hereafter, Giorgi's method (Giorgi et al., 2017). This method centers the analysis on the meaning of the experience and aims to describe its structure by identifying central themes (Giorgi et al., 2017). In this sense, the psychological structure of the experience refers to how the subject makes sense of their own lived experience in the world. Further complementing this, we analyzed a diachronic structure, aligning the phenomenological categories with each temporal phase in line with the recommendations from the microphenomenological method (Valenzuela-Moguillansky and Vásquez-Rosati, 2019). To enhance the integrity and consistency of our phenomenological exploration, we adopted an iterative method within the triangulation process, including an inter-rater agreement index (Martínez-Pernía et al., 2023). The details of this comprehensive process will be outlined below (Figure 1).

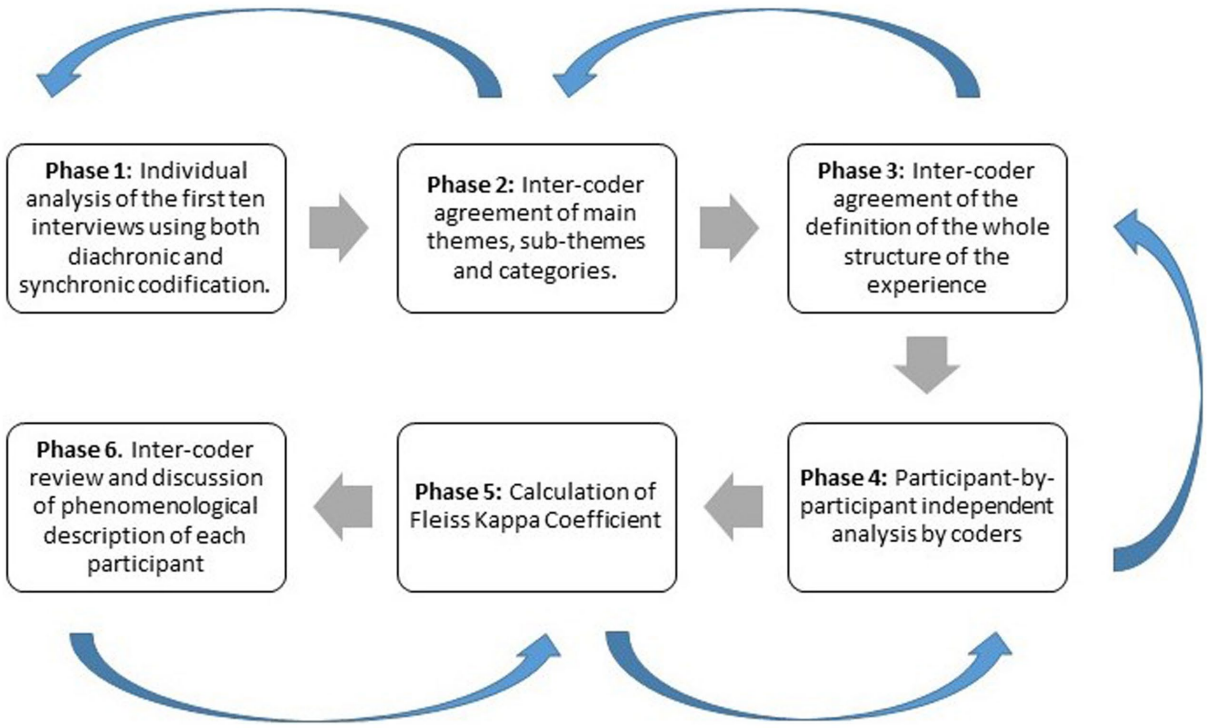


FIGURE 1
Intersubjective validation analysis. The blue arrows show the iterative process. The analysis starts with an independent codification of a part of the total research sample by each researcher. Phases 2 and 3 show the agreement of the formation and grouping of phenomenological categories. Once the emerging categories have been defined in Phase 4, the sample from Phase 1 and the total sample are analyzed again. If new categories appear, processes 2 and 3 are repeated (iterative process). In phase 5, the agreement is calculated using Fleiss' Kappa. In phase 6, the experience structure and categories of each subject are reviewed among researchers. Only discrepancies between researchers and their sources are discussed. In case of doubts in the categories (definition and grouping), we return to the previous processes and reanalyze the subject.

Initially, and supported by Atlas.ti 9 qualitative data analysis software, each of the three researchers involved in the analysis (DMP, AT, KB) began with a thorough reading of the interview and extracting a concise overview to capture the essence of each participant's entire described situation. After this initial step, the researcher underscored statements directly reflecting the experience and termed them 'meaning units'. Throughout this process, the researchers transformed participants' expressions into diachronic (elements of the experience evolving) and synchronic (elements of the experience occurring at a specific moment) codes that emphasized their inherent psychological significance. These codes were then specifically defined based on the first ten interviews (phase 1).

Following the initial coding, the researchers engaged in a process of abstraction, refining the detailed phenomenological codes such as "body tension" and "body anguish" into more abstract categories like "body feeling." This was done by identifying shared experiences across participants based on the first ten interviews (phase 2). As primary themes (e.g., "bodily resonance") and their associated subthemes (e.g., "affective quality") emerged, the research team discussed them thoroughly. Through these deliberations, they achieved a unified understanding of the main themes, subthemes, categories, and temporal phases, thus ensuring the reliability and validity of the findings.

The next step (phase 3) involved capturing and describing the overall structure of the experience based on the analysis conducted in the first ten interviews. Because the themes provide insights into specific aspects of the experience, we integrated them to form a comprehensive structure. This structure progressed from particular elements to the participants' fundamental comprehension, which was achieved by examining and systematically varying these specific elements to uncover their essence. While the initial abstraction process was conducted based on the first ten interviews until saturation was achieved, the refinement process continued throughout all subsequent interviews and phases.

Next, in phase 4, the researchers conducted, independently, a subject-by-subject analysis of all samples using the main themes, sub-themes, and categories labeled in phase 3. Subsequently, in phases 5 and 6, a novel inter-coder triangulation approach was employed to address potential errors, omissions, doubts, and disagreements among the researchers.

Phase 5 entailed calculating the Fleiss Kappa coefficient. The coefficients were used throughout the analysis process to assess the level of agreement between the researchers. By calculating the level of agreement, the researchers could identify errors, omissions, and disagreements. The average of the Kappa coefficient was 0.92 (for more information see https://osf.io/dtcr2/). This value indicates a high level of consensus among coders. The feedback provided in Phase 5 further aided in pinpointing and resolving any discrepancies among the researchers in Phase 6.

In the final step (phase 6), the researchers engaged in a collaborative review and discussion of the experiential dimensions of each analyzed participant based on the level of agreement with phase 5. This collaborative process allowed them to collectively examine and resolve any uncertainties or discrepancies related to definitions, errors, or omissions. Through this rigorous exchange of ideas, a consensus

was reached among the researchers, thus enhancing the overall robustness and credibility of the study.

Our analysis included a process of iteration throughout the procedure. If during an interview analysis, a new theme or subtheme appeared or was modified, we reviewed all the previously analyzed data to maintain consistency among the new and previous categories. This review procedure and consistency were also implemented in the last stages.

The triangulation was performed in R statistical programming language to identify the phenomenological categories in which there was agreement or disagreement between the researchers.

2.6 Statistical analysis

Repeated-measures ANOVA with the two within-subjects factors (time windows and condition.) was used for each variable (postural movement, electrodermal activity and heart rate). Greenhouse–Geisser corrections were used when the assumption of homogeneity of covariances was violated (as determined by Mauchly tests of sphericity). Bonferroni-corrected *post hoc* pairwise comparisons were computed to examine interactions and omnibus main effects. All the analyses were performed using R Studio. A significance level of $p = 0.05$ was used for all statistical analyses.

All the data analyses of this study are included in the following link (https://osf.io/dtcr2/).

3 Results

The experimental manipulation successfully demonstrated differences between the visualization of videos that depict situations triggering empathy for pain (videos related to falls of sportspeople) and the baseline condition (neutral videos related to domestic spaces). This was indicated by significantly higher arousal levels in the pain condition (mean = 6.6 ± 1.33) compared to the baseline condition (mean = 2.46 ± 1.17, $p < 0.001$), along with notably lower valence in the pain condition (mean = 3.8 ± 2.07) in contrast to the baseline condition (mean = 6.4 ± 1.99, $p < 0.001$). In terms of result presentation, we will first detail the findings concerning sensorimotor and physiological responses, referred to as the third-person results (For a detailed examination of the data refer to open data https://osf.io/dtcr2/). Subsequently, we will delve into the phenomenological outcomes or the first-person results. Finally, we will provide an integrated view combining both the third-person and first-person findings, referred to as the "bodyssence result."

3.1 Third-person results

3.1.1 Postural movement

To compare the second-to-second temporal fluctuations in participants' postural movement responses between the visualization of videos related to pain (empathy for pain condition) and neutral videos (baseline condition), an ANOVA with the factors Condition (2) and Time Window (7) was conducted. The ANOVA showed a significant interaction ($F_{6,204} = 13.25$, $p < 0.001$) between emotional conditions and the time window in the anteroposterior postural movement. The main effect of time windows was not statistically significant ($F_{6,204} = 1.94$, $p = 0.076$). There was a significant main effect of condition ($F_{6,204} = 18.32$, $p < 0.001$). As shown in Figure 2, there were significant differences in the time windows [−3], [2], and [3] ($p < 0.05$) between conditions. In the empathy for pain condition, there was a significant increase ($p < 0.05$) in the post-fall time windows when compared to the initial windows in specific time windows (for details see Figure 2). Interestingly, the last post-fall windows show significantly less postural movement than the previous one. In the baseline condition, there was a statistically significant decrease ($p < 0.05$) when comparing the initial windows to the final windows.

3.1.2 Electrodermal activity

To compare the temporal fluctuations in electrodermal activity between the empathy for pain condition and the baseline condition, an ANOVA with the factors Condition (2) and Time Window (7) was performed. The ANOVA showed a significant interaction ($F_{6,204} = 3.04$, $p = 0.007$) between emotional conditions and the time window in electrodermal activity. There was a significant main effect of time windows ($F_{6,204} = 2.6$, $p = 0.019$) and a significant main effect of condition ($F_{5,204} = 15.24$, $p < 0.001$). As shown in Figure 3, the empathy for pain condition reported significantly more phasic activity than the baseline condition in each temporal window ($p \leq 0.01$). No significant differences were found in the pairwise comparisons between temporal windows at each condition.

3.1.3 Heart rate

To compare the temporal fluctuations in heart rate responses between the empathy for pain condition and the baseline condition, an ANOVA with the factors Condition (2) and Time Window (7) was conducted. The ANOVA found no significant interaction ($F_{6,204} = 1.32$, $p = 0.25$) in the heart rate. There was a main effect in the time window in the empathy for pain conditions ($F_{6,204} = 4.79$, $p < 0.001$). There was no significant main effect of emotional condition. As can be seen from Figure 4, empathy for pain conditions reported a significant decrease ($p < 0.05$) in the last two windows [3, 4] compared to the window before the fall [−1].

3.2 First-person result

This section presents the main findings of the phenomenological analysis conducted on the microphenomenological interviews following the participants' viewing of fall videos. The analysis focused on the temporal dynamics of the experience. First, we present descriptions of the phenomenological dimensions that emerged in the analysis, followed by the full temporal structure of the empathy experience (a more comprehensive analysis of 28 out of the 35 participants, as well as a review of the codebook, can be found in our previous publication (Martinez-Pernía et al., 2023).

3.2.1 Phenomenological categories

Following our analysis, four main phenomenological categories emerged (temporality, bodily resonance, kinesthetic motivation, and attentional focus). One category is a temporal category that persists throughout the entire empathic experience, evolving in synchrony with the events sportspersons undergo while engaging in extreme

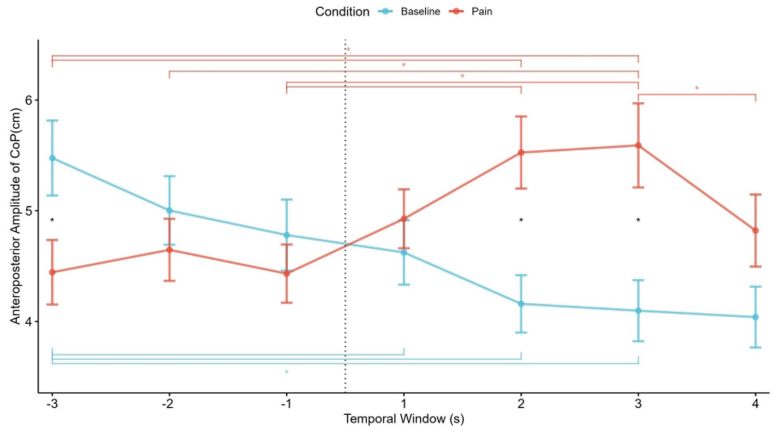


FIGURE 2
Temporal fluctuations of mean anteroposterior postural movement responses in participants during the visualization of pain-related videos (empathy for pain condition) and neutral videos. The dashed line between [−1] and [1] temporal window represents the fall of the sportsperson. The asterisk (*) denotes significant differences between conditions (black) and between temporal windows in empathy for pain (red) and baseline (cyan).

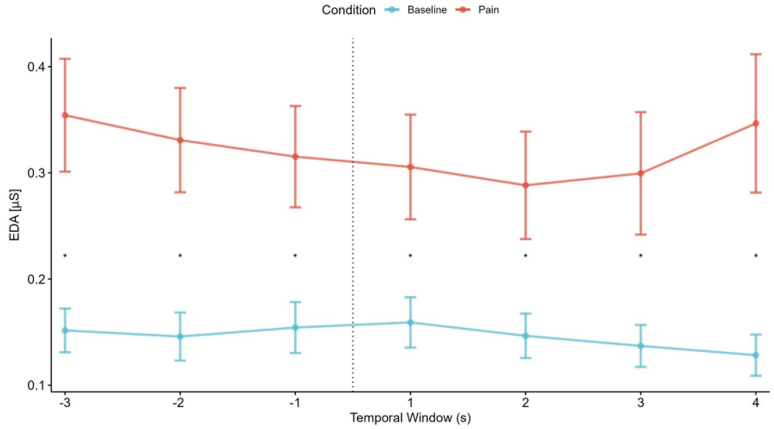


FIGURE 3
Temporal fluctuation in participants' electrodermal activity responses to pain-related videos (empathy for pain condition) and neutral videos. The figure shows the temporal fluctuation of mean electrodermal activity. The dashed line between [−1] and [1] temporal window represents the fall of the sportsperson. The asterisk (*) denotes significant differences between conditions (black).

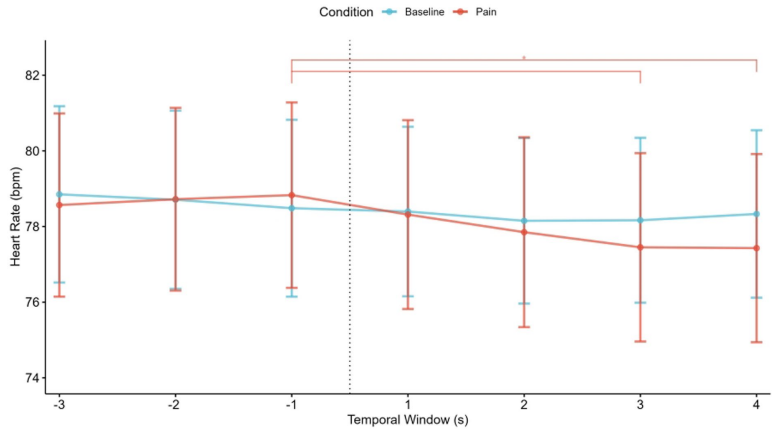


FIGURE 4
Temporal fluctuations in participants' heart rate responses to pain-related videos (empathy for pain condition) and neutral videos. The figure shows the temporal fluctuation of mean heart rate. The dashed line between [−1] and [1] temporal window represents the fall of the sportsperson. The asterisk (*) denotes significant differences between temporal windows in empathy for pain (red).

sports. The other three categories, which traverse temporality, are termed bodily resonance, kinesthetic motivation, and attentional focus. Below, we delve into these dimensions of the experience.

A consistent temporal structure was identified in the analysis and was observed across all participants. This temporal pattern exhibits fluctuations that occur in synchrony with the events experienced by the sportsperson. These fluctuations are framed within three distinct experiential phases: anticipatory, climax, and recovery. During the anticipation phase, which develops before the sportsperson falls, the participants intuitively perceive the forthcoming occurrence of an accident, marked by a gradual intensification of bodily sensations.

> *"It's going up, I feel a nerve… as if one were seeing it in real life… it's a sensation, I could call it an instinct that it's going to fall… it's going to fall…"* (P20).

During the climax phase, when the sportsperson is about to fall and immediately during the fall, participants experienced the maximum intensity of bodily feelings. An excerpt from a participant's interview is provided below.

> *"I feel the tension as if it were contracting, releasing, here it enters my stomach; there I feel the, the negative. I also feel my whole body more, more alert in front of these videos. In the moments, of the fall itself, of when "paff" the person hits the ground, there I felt it stronger."* (*P10*).
>
> *Finally, in the recovery phase, which corresponds to the moments after the sportsperson's fall, participants experienced a decreased intensity of bodily feelings*
>
> *"…the sensation is brutally relieved when it already fell"* (*P12*).

In the analysis of synchronic categories within each temporal phase, three primary categories were identified: bodily resonance, kinesthetic motivation, and attentional focus.

The first category, called bodily resonance, captures the intimate, lived experience of participants as their bodies are affected by and moved by the actions of the sportspersons. For example:

> *"I know a fall is coming, like first being expectant before the fall. I feel a little nauseous, anguish too, and my chest is tight. Being with this feeling of wanting to get away, like feeling that my body was going backward".* (P10).

The second category, called kinesthetic motivation, captures the driving force experienced by individuals when witnessing the actions of sportspersons, prompting them to either safeguard themselves (protective motivation) or assist others (prosocial motivation).

An example of protective motivation is as follows:

> "*…is the sensation of the body that you clench everything: your hands, arms, legs and back muscles tighten. It is like a feeling of protection, of how the body feels threatened by a predator, so to speak, that is the sensation of activating all the senses and tightening up".* (*P26*).

An example of prosocial motivation is as follows:

> *"My muscles were contracting more, I felt my knees buckling. I felt like, if I moved somehow, I was going to keep them from falling. I kind of engaged with them and felt that if I controlled my body, they wouldn't fall"* (*P16*).

And finally, the third phenomenological category named attentional focus, reflects the very spots and instances that captivate and hold participants' attention as they watch sportspersons in action during extreme sports. For example:

> *"I immediately noticed that my body was very different… I felt my muscles contracting"* (*P20*).

3.2.2 Experience structure

The previously mentioned four main phenomenological categories (temporality, bodily resonance, kinesthetic motivation, and attentional focus) are strongly intertwined in an empathic experiential structure. For a complete visualization of the empathic structure and a more comprehensive breakdown, including the percentages of participants experiencing each aspect, please refer to Figure 5.

3.3 Anticipatory phase

During the anticipatory phase of bodily resonance, participants engaged in a profound bodily experience triggered by the physical actions of the sportspersons. This phase was primarily characterized by a keen sense of anticipation as participants intuitively grasped the unfolding events and directed their attention toward the future state of the sportsperson. Notably, a significant majority of participants described multifocal sensations, underscoring the widespread distribution of bodily resonance throughout their physical beings. The localization of these sensations were experienced in prominent bodily regions including the abdomen, chest, heart, face, and extremities. These sensations were intertwined with a subtle yet discernible undercurrent of negative affective valence, gradually intensifying as the experience transitioned toward the climactic phase. This negative valence was manifested through various affective qualities such as tension, pain, fear, anguish, and anxiety, with tension as the predominant emotional tone. Additionally, participants experienced subtle kinesthetic sensations, which manifested as sporadic instances of increased bodily movement and transient feelings of imbalance. Furthermore, they navigated the interplay between two distinct motivations—a self-protective motivation and a prosocial motivation. Participants' self-protective motivation was marked by a sense of detachment while watching the video. It was as though participants' bodies were preparing to distance themselves from the impending event. Simultaneously, their attention was drawn inwards, focusing on personal discomfort and sensations of rejection. In contrast, participants' prosocial motivation was driven by a desire to prevent injury or a fall in the other person and their attention directed toward others. This outward attention encompassed feelings, actions, or thoughts directed at the suffering individuals, revealing the complexity of their responses during this phase of bodily resonance.

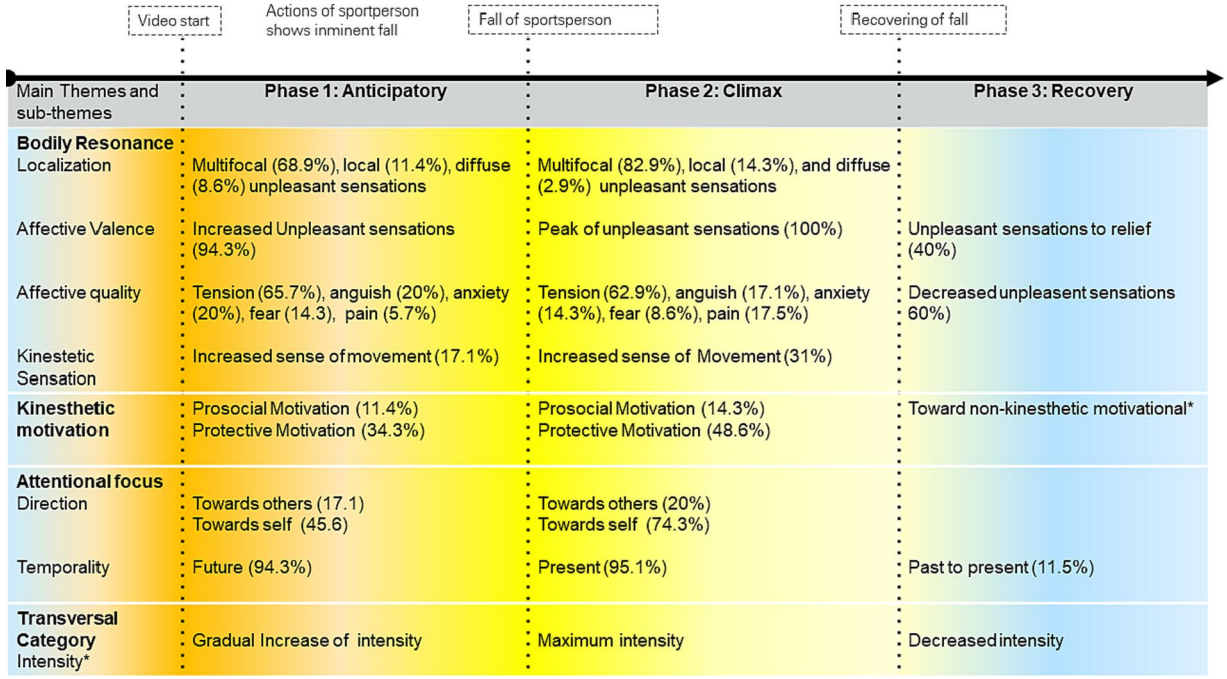


FIGURE 5
Temporal structure of empathy for pain. The arrow at the top of the figure indicates the time evolution of the experience. The dotted lines show the points of change from one phase to the next phase. Above the arrow are shown the video's events concerning the participants in each phase. The cross intensity refers to the category that characterizes each phase. This category is integrated with the video moments perceived by the participants to define each phase. The asterisk (*) indicates an emergent category following refinement of the temporal structure, previously described as none. The color shows the temporal flow of intensity from gradual increase (orange), to climax (yellow), and then decrease (blue). The % represents the percentage of the total sample in the category.

3.4 Climax phase

In the climax phase of bodily resonance, participants continued to undergo significant changes in their bodily experiences and emotional responses concerning the fall of the sportsperson. The participants' attentional focus during this phase shifted toward a synchronized experience with the suffering of others in the present moment of fall. This shift from future anticipation marked a temporal synchronization of sensations concerning the observed pain, intensifying their connection to the sportsperson's experience. Importantly, negative affective valence reached its peak intensity during this phase as participants experienced a heightened emotional response while observing the sportsperson's pain. The localization of bodily sensations shifted to prominent areas such as the abdomen, chest, heart, face, and extremities, reflecting a different pattern from the anticipatory phase. Affective quality remained characterized by sensations of tension, but pain sensations became more pronounced.

Kinesthetic sensations, such as increased movement and feelings of bodily imbalance were more prominent during the climax phase, indicating a greater level of subjective engagement. This phase represented the zenith of bodily resonance, marked by heightened emotional intensity and kinesthetic involvement in response to the observed pain. Furthermore, during this phase, the experience of kinesthetic motivations was heightened. For instance, self-protective motivation remained prevalent, accompanied by sensations of heightened bodily tension and a sense of detachment from the other person's suffering. Prosocial motivation persisted, reflecting the desire to prevent injury or a fall concerning the other person.

Interestingly, both in the preceding phase and the current one, participants identified two distinct experiences, termed self-centered empathy and other-centered empathy. Self-centered empathy is characterized by an attentional focus on one's own experience, driven by a motivation for self-protection (N = 27). On the other hand, other-centered empathy involves directing attention toward the experience of the sportsperson's pain and is accompanied by a motivation to offer assistance ($N = 8$).

3.5 Recovery phase

The recovery phase marked a transition in bodily resonance, reflecting a state of relief following the intense experiences of the climax phase. In this moment, participants' attention was enveloped by a palpable sense of relief, and, some were attuned to the aftermath of the fall. During this phase, participants reported a decrease in the intensity of negative affect, signifying a release from the heightened emotional state observed in the climax phase. The affective quality transitioned toward sensations of relief as participants began to feel a tangible release from the preceding emotional intensity. In contrast to earlier phases, participants felt kinesthetic sensations and motivation diminish during the recovery phase. Overall, the recovery phase represented a shift toward emotional relief and a return to a more stable bodily state following the intense emotional and kinesthetic responses observed in the climax phase. Figure 5 shows the temporal structure of empathy for pain.

3.6 Bodyssence result

After presenting the main results from physiological and phenomenological data, this section provides an integration of both dimensions. The phenomenological findings highlight three temporal phases: anticipatory, climax, and recovery. In the neurophysiological data, we identify three sequential temporal phases: the pre-fall period followed by the post-fall period, which is further divided into an early and a later response. Integrating both sets of data (1p and 3p) gives rise to three synchronously interwoven phases: the anticipatory temporal phase coinciding with the pre-fall period, the climax temporal phase aligning with the early post-fall response, and the recovery temporal phase paired with the later post-fall response. Thus, we illuminate the 'bodyssence of empathy,' a fusion of third-person neurophysiological data and first-person phenomenological insights. At its core, bodyssence exemplifies the rich interplay between our physiological responses and our deeply felt experiences coordinated with the events occurring in the environment (Figure 6). The three successive and distinctive phases of "empathy bodyssence" are listed and described as follows:

Forefeel: When extreme sportspersons began performing their acrobatics, participants experienced a pre-reflective felt knowledge or prediction of an accident. They experienced a gradual increase in the intensity of their negative emotions, multiple bodily sensations, low postural movement (AP-COP), and a high heart rate during the temporal dynamic of empathy. The longer participants watched the sportsperson, the more intense their experience. In this phase, participants also felt different kinesthesis motivations and focused attention. The increase in emotional intensity was accompanied by a high electrodermal activity compared to baseline condition.

Fullfeel: At some point during the sportsperson's acrobatics, they lost their balance; the impact with the ground lasted a few brief seconds. In these moments, participants' bodily resonance peaked. They were immersed in intense negative emotions, particularly in response to the sportsperson's evident pain. This emotional surge was mirrored in bodily sensations, most prominently in the abdomen, chest, heart, face, and extremities, intertwined with kinesthetic sensations of movement and imbalance. As these feelings escalated, the amplitude of postural movement increased and reached its maximum point in seconds 2 and 3 post-fall, while the heart rate decreased in its last second, and the electrodermal activity remained consistently high.

Reliefeel: As the sportsperson was attempting to rise from the ground after their fall, there was a shift in the participants' experiential state. Their previously heightened emotional intensity began to wane, mirrored by the physical relaxation they felt coursing through their bodies. This easing of negative emotions manifested in sensations of relief, tranquility, and diminished concern. Concurrently, as the participants watched the sportsperson try to recover, their postural stability began to be restored in the last second, evidenced by a reduction in the amplitude of their anteroposterior movements. This sensorimotor return to equilibrium was characterized by the sustained low heart rate. The electrodermal activity remained high compared to baseline condition.

4 Discussion

Our research endeavored to explore the multifaceted nature of empathy within an enactive framework. To facilitate this exploration, we introduce the novel concept of "bodyssence" (body + essence). This neologism encapsulates the holistic convergence of our neurophysiological reactions with our lived, subjective experiences. Our primary objective was to study how the temporal evolution of motor, physiological, and phenomenological facets shapes our empathetic consciousness in synchrony with the experience of the sportsperson enduring a physical accident. Our research on empathy bodyssence reveals a bodily experience marked by three consecutive phases named Forefeel, Fullfeel and Reliefeel. Bodyssence emphasizes the dynamic and temporal nature of these dimensions offering insights into how our embodied experiences evolve and resonate with both our internal states and external influences.

In the following paragraphs, we explore two significant concepts. These are the temporal dimension of bodyssence within the overall experience and the pivotal role that bodily resonance plays in fostering empathy for pain.

4.1 The temporality of bodyssence

The temporality of the bodyssence concerning the external referent of the fall suggests that there is a bodily attunement of the observer with the expressive behavior of the sportsperson shown in the video. Body attunement has been reported in several studies that show temporal coordination of individuals' behaviors that manifest spontaneously in our daily interactions. For example, studies have identified that when observing the pain of another, there is a physiological and brain synchrony (Goldstein et al., 2017, 2018; Peled-Avron et al., 2018). In our study, the temporal coordination of bodyssence has been found in the three phases, Forefeel, Fullfeel, and Reliefeel.

In the Forefeel phase, the temporality of the embodiment reveals that when participants direct their attention toward an impending future fall, they experience multiple bodily sensations and show less postural sway. Similar to the findings of our study, previous studies have demonstrated that by perceiving the bodily cues of another, agents can predict the future actions of the another (Iacoboni et al., 2005). Likewise, the bodily cues of another being are perceived as a meaningful action and affective state through bodily resonance (Tanaka, 2015). Overall, the body's feeling and movement response shows that the bodyssence is an active participant in the implicit knowledge of the future state of another (the oncoming fall).

In the Fullfeel Phase, the findings reveal that in the maximum intensity of the videos, the embodiment appears with a higher postural oscillation, bodily sensation, and unpleasant valence. This finding is consistent with the findings of studies previous that illustrated the responses of the sensory-motor cortex (Riečanský and Lamm, 2019), postural movement (Gea et al., 2014; Lelard et al., 2017, 2019), autonomic system (Eisenberg et al., 1988, 1991) and phenomenological insights (Fuchs, 2017) in the presence of others' suffering. Despite the constant electrodermal activity in all phases, its elevation in the empathy for pain condition indicates autonomic embodied resonance to others' pain, consistent with previous studies (Hein et al., 2011; Lelard et al., 2014).

In addition, subjects observing the fall report a set of affective qualities and different bodily sensations in different parts of the body. This reflects that a bodily resonance, not only occurs in objective bodily phenomena but can also be accessed from subjectivity. Bodily

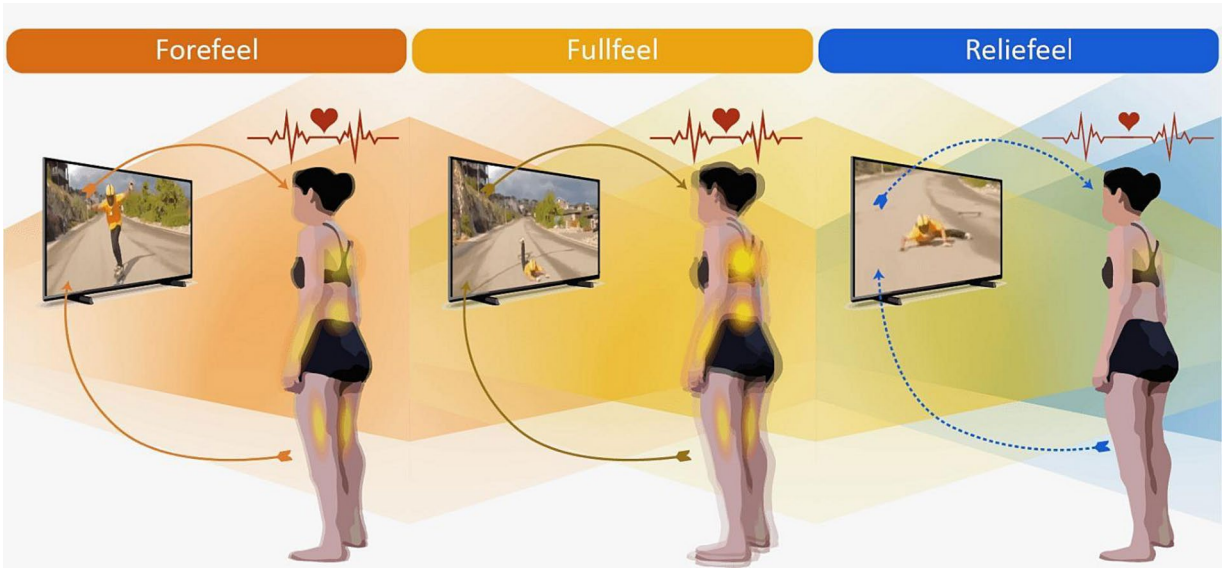


FIGURE 6
Illustration of Empathy Bodyssence. This figure depicts the interwoven phenomenological and neurophysiological phases of empathy for pain, graphically representing the seamless integration of first-person experiential data (1p) and third-person physiological data (3p). These distinct phases collectively illuminate the 'bodyssence of empathy,' showcasing the dynamic interplay between an individual's physiological reactions and their concurrent lived experiences in response to observed events. Forefeel (orange hue), as the initial phase, embodies the body's intuitive pre-reflective knowledge, with participants sensing the imminent physical accidents of extreme sportspersons, a state that is characterized by the most subdued levels of postural movement and elevated heart rate during the temporal dynamic of empathy. Fullfeel (yellow hue) represents the peak of empathetic connection, characterized by intense negative emotions, prominent bodily sensations, heightened kinesthetic sensations. This phase is distinguished by an increase in postural movement and decrease in heart rate in specific moments of the temporal dynamic of empathy. Reliefeel (blue hue): participants experienced a decrease in emotional intensity and a concurrent sense of relief, with postural control returning to an equilibrium and heart rate remaining low. Throughout all three phases, electrodermal activity was consistently elevated compared to the baseline condition.

resonance also includes the tendency to action or kinesthetic motivation, which do not necessarily manifest themselves in physical space, but are phenomena that manifest themselves in lived space (Fuchs, 2017).

In the Reliefeel Phase, temporal attunement has also been revealed with the decrease of intensity of the bodily feeling experience and the decrease of postural sway and heart rate (Lang et al., 1993; Lelard et al., 2014, 2017; Mouras and Lelard, 2021). While some studies interpret the reduction in postural sway and heart rate after exposure to highly aversive film content, such as scenes depicting car accidents and dead bodies, as indicative of a freezing-like response (Hagenaars et al., 2014), another proposal suggests that rapid deceleration in heart rate following a stressful situation allows a recovery state in the absence of imminent danger (Bernston et al., 1993). This implies, in conjunction with phenomenological data, that the decline in heart rate may be linked to a diminution of anguish following the peak experience of witnessing someone else fall.

Interestingly, the heart rate demonstrates swift adaptation and decreases, while electrodermal activity remains relatively unchanged. These findings suggest that, even though sympathetic system activation persists, as indicated by elevated electrodermal activity, subtle changes such as an increase in the parasympathetic system or a decrease in the sympathetic system (e.g., Weissman and Mendes, 2021) could signify an augmented state of relaxation following the observation of the fall. Additionally, the temporal invariability of the electrodermal activity aligns with previous reports indicating a slow temporal adaptation of electrodermal activity to exposure to aversive stimulus (Lelard et al., 2014).

The depiction of these phases reveals a noteworthy alignment between subjective experiences and physiological data, which has been a subject of controversy in some studies on empathy for pain. For instance, a study shown that mental simulation had a modulatory effect on postural sway but not on self-reported measures of empathy (Beaumont et al., 2021). While the authors debate the underlying physiological mechanisms contributing to the absence of results in self-reporting, it seems that enhancing subjective data collection could yield more detailed insights about the findings (Martínez-Pernía et al., 2021; Vergara et al., 2022). Our study underscores the importance of deepening subjective experiences to accurately grasp how individuals experience specific points of exposure to the video. This allows us to make a more detailed interpretation of the physiological findings. Thus, future research adopting traditional paradigms could benefit from incorporating phenomenological interviews or even simpler phenomenological self-reports, as previously utilized in research (Grice-Jackson et al., 2017a,b).

4.2 Beyond freezing and fight or flight actions: the body as a source of primary empathy

Our findings concur with a multitude of studies emphasizing the pivotal role the body plays in empathetic responses, particularly in the context of pain. For instance, recent research has highlighted that specific neural regions activated during first-hand pain experiences are similarly triggered when observing the pain of others (Riečanský

and Lamm, 2019). Specifically, when individuals witness somatic pain, such as viewing others in painful situations, there occurs a pronounced activation in regions associated with negative emotional processing and in areas related to somatosensation and skeletal muscle control (Riečanský and Lamm, 2019). Studies suggest that emotion recognition becomes compromised when facial mimicry is restricted, whether through the impossibility of using certain muscles (e.g., biting a pen while observing others' emotions) (Oberman et al., 2007; Borgomaneri et al., 2020) or muscle immobilization via botox injection (Neal and Chartrand, 2011), as well as when bodily movement is restricted (Reed et al., 2020). Collectively, this body of evidence underscores the indispensable role of bodily experiences in comprehending the emotions of others.

Considering the role of the body in social cognition, objective bodily responses and subjective bodily descriptions go beyond the consideration of mere freezing or flight responses to an aversive stimulus (Azevedo et al., 2005; Hagenaars et al., 2014). Rather, we suggest that bodily experience has an interpersonal function to allow for resonance with and understanding of the other person. This means that the other's bodily expression appears to us as meaningful and affective actions that express their intentions, needs, and goals in a shared context (Gallagher, 2012; Tanaka, 2015; Fuchs, 2017). Those expressions affect us and are experienced through and with our bodies (Cea and Martínez-Pernía, 2023), which defines the concept of bodily resonance. Fuchs (2013, 2017) has proposed two components of bodily resonance: an affective dimension (the body is affected by events through bodily sensations) and an e-motive dimension (the body tends to act through body movement). Both components are related to our findings that show an affective quality of bodily sensations, kinesthetic sensations and motivation, and changes in postural sway.

Within the emotional dimension, participants expressed dual motivations: to protect themselves and/or to adopt the perspective of the person experiencing distress and assist the person suffering the fall. These intentions highlight the active facet of bodily resonance and suggest that it also modifies bodily sensations in a nuanced sensorimotor cycle. Intriguingly, past research has recognized both aversion-motivated and approach-motivated states. In our findings, we discerned two distinct experiences among participants—those driven by self-oriented protective actions and those motivated by other-oriented actions. Accompanying these intentions to act were differing attentional foci—some directed inwardly and others directed outwardly. Both dimensions underscore the intricacies of these phenomena (for more details, see Martinez-Pernía et al., 2023) that previous studies, which largely relied on physiological reports and self-reporting, might not have fully captured. Further research could explore the relationship between both experiences and physiological responses. For this purpose, a special focus might be needed to increase the sample size and conduct robust between-group comparisons. For example, our study lacks sufficient statistical power to differentiate between these two experiences, given the unbalanced sample sizes in each group (27 self-protective vs. 8 prosocial). Incorporating such insights into subsequent investigations will undoubtedly contribute to a more comprehensive understanding of these relationships.

In another vein, the results of this study show that bodily sensations were described in a multifocal manner during the climax in various parts of the body, encompassing the level of the chest and/or extremities. No subject described a general pain sensation or an absence of bodily sensation. This contrasts with what was found by Grice-Jackson et al. (2017a), who found three different experiential structures in a classical picture-based paradigm: non-responder, sensory/localizer, and affective general. We suggest that given the intensity and type of stimulus, our experiential structures cluster into multifocal responders.

4.3 Limitations of the study

One of the limitations of the study is the absence of an ecological stimulus that occurs in natural contexts. Therefore, the conclusions raised in this article are limited to the laboratory context. The authors propose to change the classical neuroscience paradigm to paradigms that have a greater natural context (i.e., with multisensory contexts, free movement, and real interaction with others; Troncoso et al., 2023). However, our study advances the ecology by allowing participants to move freely in a bipedal position. The use of the instruments enabled us to evaluate embodiment phenomena by allowing free movement in a natural bipedal position. It is suggested that future empathy studies use paradigms in natural contexts and instruments that allow the evaluation of neural correlates at the same time (e.g., mobile eeg).

Another limitation is that we queried subjects about the impact of the maximum intensity video, which was different for everyone and could generate intragroup differences in the type of behavior. However, in the interviews, the subjects explicitly stated that they presented a similar experience in the other videos. For future studies, we suggest using the same scenes for all participants and integrating the same physiologically evaluated scenes.

Another limitation of this study pertains to the baseline condition video, which inherently differs from the videos showing the falls. Traditionally, the empathy for pain paradigm utilizes more analogous images—one showing pain and the other not showing pain. However, in our study, the baseline condition was specifically designed to gauge a fundamental neurophysiological and postural response. This provides a foundational reference allowing us to effectively compare the heightened responses evoked by the empathy for pain condition. However, an important advantage of our study is the integration of the phenomenological approach. While our stimulus may not be a direct control condition in comparison to the empathy for pain video, our experimental design rooted in phenomenology allows us to delve deeply into the moment-to-moment lived experiences of participants. By doing so, we not only understand the genuine reactions of participants but also grasp the intricate dynamics of empathy as experienced in real-world conditions. For future research, we suggest considering the use of comparable videos instead of videos with different content to enhance experimental control. This would ensure that participants' responses are more comparable and that differences in stimulus nature do not become a source of confusion. Improving the uniformity of stimuli can contribute to greater robustness and generalizability of findings.

Additionally, expanding the research scope could involve extending physiological data collection before and after stimuli exposure, along with deepening phenomenological interviews. This would offer a comprehensive understanding of physiological responses, clarifying baseline patterns and addressing gaps like

those seen in electrodermal activity analysis. Additionally, exploring anticipatory physiological changes before stimuli exposure and post-stimulus recovery could unveil preparatory mechanisms and adaptive responses. Such an approach is especially promising for dynamic stimuli studies, offering more insights into temporal dynamics.

4.4 Constraints on generalizability

The sample was confined to a segment of young, educated, and healthy individuals, which may not be representative of the broader population. It is also crucial to consider potential cultural and contextual factors that could influence the phenomenological and physiological aspects of empathy. For instance, individuals consistently exposed to pain situations, such as health professionals (Decety et al., 2010), and those displaying lower levels of empathic resonance, like individuals with psychopathy (Decety et al., 2013), may find it worthwhile to replicate this study. Understanding how individuals from different backgrounds perceive the suffering of others could offer valuable insights into the interplay between experience-physiology and context.

5 Conclusion

Our results reveal a temporal structure of the bodily experience and physiology of empathy for pain, marked by three consecutive phases that synchronize with the evolving experience of the sportsperson. This highlights how the body resonates dynamically, both in its subjective realm and physiological realm, when experiencing the suffering of another. Furthermore, the current enactive framework illustrates how the mutual interaction between phenomenological data and physiological data allows them to inform each other with higher precision to understand the empathy bodyssence.

This research not only validates the depth and applicability of the enactive approach but also encourages a collaborative effort among philosophers and neuroscientists, and associated disciplines. The development of empirical evidence within the enactive framework could not only refine the theoretical propositions but also, and only in this way, ensure the paradigm's validity and enduring significance, thereby paving the way for its lasting legacy.

Data availability statement

The datasets and analysis presented in this study can be found in online OSF repository: https://osf.io/dtcr2/.

Ethics statement

The studies involving humans were approved by the "Scientific Ethics Committee of the Servicio de Salud Metropolitano Oriente" and the "Research in Humans being Ethics Committee of the Medicine Faculty, Universidad de Chile." The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AT: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. DM-P: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing. KB: Data curation, Formal analysis, Methodology, Writing – review & editing. ÁR-R: Data curation, Formal analysis, Methodology, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. AT was supported by the ANID grant # 21220194. DM-P was partially supported by the ANID/FONDECYT iniciación 11190507 and ANID/FONDECYT regular 1241087. KB was supported by the ANID grant # 21232191.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Azevedo, T. M., Volchan, E., Imbiriba, L. A., Rodrigues, E. C., Oliveira, J. M., Oliveira, L. F., et al. (2005). A freezing-like posture to pictures of mutilation. *Psychophysiology* 42, 255–260. doi: 10.1111/j.1469-8986.2005.00287.x

Beaumont, A., Granon, S., Godefroy, O., Lelard, T., and Mouras, H. (2021). Postural correlates of painful stimuli exposure: impact of mental simulation processes and painlevel of the stimuli. *Exp. Brain Res.* 239, 1929–1936. doi: 10.1007/s00221-021-06102-y

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. *San Antonio, TX: Psychological Corporation*. doi: 10.1007/978-14419-1005-9_441

Bernston, G. G., Cacioppo, J. T., and Quigley, K. S. (1993). Respiratory sinus arrhythmia: autonomic origins, physiological mechanisms, and psychophysiological implications. *Psychophysiology* 30, 183–196. doi: 10.1111/j.1469-8986.1993.tb01731.x

Borgomaneri, S., Bolloni, C., Sessa, P., and Avenanti, A. (2020). Blocking facial mimicry affects recognition of facial and body expressions. *PLoS One* 15:e0229364. doi: 10.1371/journal.pone.0229364

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* 25, 49–59. doi: 10.1016/0005-7916(94)90063-9

Bufalari, I., Aprile, T., Avenanti, A., Di Russo, F., and Aglioti, S. M. (2007). Empathy for pain and touch in the human somatosensory cortex. *Cereb. Cortex* 17, 2553–2561. doi: 10.1093/cercor/bhl161

Cea, I., and Martínez-Pernía, D. (2023). Continuous organismic sentience as a synthesis of Core affect and vitality. *J. Conscious. Stud.* 30, 7–33. doi: 10.53765/20512201.30.3.007

Colombetti, G. (2014). *The feeling body: Affective science meets the enactive mind*. Massachusetts: MIT Press.

De Jaegher, H., and Di Paolo, E. (2007). Participatory sense-making: an enactive approach to social cognition. *Phenomenol. Cogn. Sci.* 6, 485–507. doi: 10.1007/s11097-007-9076-9

de Waal, F. B. M., and Preston, S. D. (2017). Mammalian empathy: behavioural manifestations and neural basis. *Nat. Rev. Neurosci.* 18, 498–509. doi: 10.1038/nrn.2017.72

Decety, J., Chen, C., Harenski, C., and Kiehl, K. A. (2013). An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Front. Hum. Neurosci.* 7:489. doi: 10.3389/fnhum.2013.00489

Decety, J., and Jackson, P. L. (2004). The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* 3, 71–100. doi: 10.1177/1534582304267187

Decety, J., Yang, C.-Y., and Cheng, Y. (2010). Physicians down-regulate their pain empathy response: an event-related brain potential study. *NeuroImage* 50, 1676–1682. doi: 10.1016/j.neuroimage.2010.01.025

Eisenberg, N., Fabes, R. A., Bustamante, D., Mathy, R. M., Miller, P. A., and Lindholm, E. (1988). Differentiation of vicariously induced emotional reactions in children. *Dev. Psychol.* 24, 237–246. doi: 10.1037/0012-1649.24.2.237

Eisenberg, N., Fabes, R. A., Schaller, M., Miller, P., Carlo, G., Poulin, R., et al. (1991). Personality and socialization correlates of vicarious emotional responding. *J. Pers. Soc. Psychol.* 61, 459–470. doi: 10.1037/0022-3514.61.3.459

Eklund, J. H., and Meranius, M. S. (2021). Toward a consensus on the nature of empathy: a review of reviews. *Patient Educ. Counsel.* 104, 300–307. doi: 10.1016/j.pec.2020.08.022

Fuchs, T. (2013). "The phenomenology of affectivity" in *The Oxford handbook of philosophy and psychiatry in Oxford Academic*. eds. K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler and G. Stanghelliniet al. (Oxford University Press), 612–631. doi: 10.1093/oxfordhb/9780199579563.013.0038

Fuchs, T. (2017). "Levels of empathy—primary, extended, and reiterated empathy" in *Empathy: Epistemic problems and cultural-historical perspectives of a cross-disciplinary concept*. eds. V. Lux and S. Weigel (Palgrave Macmillan/Springer Nature), 27–47.

Gallagher, S. (2003). Phenomenology and experimental design toward a phenomenologically enlightened experimental science. *J. Conscious. Stud.* 10, 85–99.

Gallagher, S. (2012). In Defense of phenomenological approaches to social cognition: interacting with the critics. *Rev. Phil. Psych.* 3, 187–212. doi: 10.1007/s13164-011-0080-1

Gallese, V. (2014). Bodily selves in relation: embodied simulation as second-person perspective on intersubjectivity. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 369:20130177. doi: 10.1098/rstb.2013.0177

Gea, J., Muñoz, M. A., Costa, I., Ciria, L. F., Miranda, J. G. V., and Montoya, P. (2014). Viewing pain and happy faces elicited similar changes in postural body sway. *PLoS One* 9:e104381. doi: 10.1371/journal.pone.0104381

Giorgi, A., Giorgi, B., and Morley, J. (2017). "The descriptive phenomenological psychological method" in *The sage handbook on qualitative research in psychology*. eds. C. Willig and W. S. Rogers (Thousand Oaks: Sage), 176–192.

Goldstein, P., Weissman-Fogel, I., Dumas, G., and Shamay-Tsoory, S. G. (2018). Brain-to-brain coupling during handholding is associated with pain reduction. *Proc. Natl. Acad. Sci. USA* 115, E2528–E2537. doi: 10.1073/pnas.1703643115

Goldstein, P., Weissman-Fogel, I., and Shamay-Tsoory, S. G. (2017). The role of touch in regulating inter-partner physiological coupling during empathy for pain. *Sci. Rep.* 7:3252. doi: 10.1038/s41598-017-03627-7

Golland, Y., Arzouan, Y., and Levit-Binnun, N. (2015). The mere co-presence: synchronization of autonomic signals and emotional responses across co-present individuals not engaged in direct interaction. *PLoS One* 10:e0125804. doi: 10.1371/journal.pone.0125804

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., and Citi, L. (2016). cvxEDA: a convex optimization approach to electrodermal activity processing. *I.E.E.E. Trans. Biomed. Eng.* 63, 797–804. doi: 10.1109/TBME.2015.2474131

Grice-Jackson, T., Critchley, H. D., Banissy, M. J., and Ward, J. (2017a). Common and distinct neural mechanisms associated with the conscious experience of vicarious pain. Cortex. *J. Devoted Study Nerv. Syst. Behav.* 94, 152–163. doi: 10.1016/j.cortex.2017.06.015

Grice-Jackson, T., Critchley, H. D., Banissy, M. J., and Ward, J. (2017b). Consciously feeling the pain of others reflects atypical functional connectivity between the pain matrix and frontal-parietal regions. *Front. Hum. Neurosci.* 11:507. doi: 10.3389/fnhum.2017.00507

Hagenaars, M. A., Roelofs, K., and Stins, J. F. (2014). Human freezing in response to affective films. *Anxiety Stress Coping* 27, 27–37. doi: 10.1080/10615806.2013.809420

Hamilton, A. K., Pernía, D. M., Puyol Wilson, C., and Carrasco Dell'Aquila, D. (2019). What makes metalheads happy? A phenomenological analysis of flow experiences in metal musicians. *Qual. Res. Psychol.* 16, 537–565. doi: 10.1080/14780887.2017.1416210

Hari, R., Himberg, T., Nummenmaa, L., Hämäläinen, M., and Parkkonen, L. (2013). Synchrony of brains and bodies during implicit interpersonal interaction. *Trends Cogn. Sci.* 17, 105–106. doi: 10.1016/j.tics.2013.01.003

Hein, G., Lamm, C., Brodbeck, C., and Singer, T. (2011). Skin conductance response to the pain of others predicts later costly helping. *PLoS One* 6:e22759. doi: 10.1371/journal.pone.0022759

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3:e79. doi: 10.1371/journal.pbio.0030079

Jauniaux, J., Tessier, M. H., Regueiro, S., Chouchou, F., Fortin-Côté, A., and Jackson, P. L. (2020). Emotion regulation of others' positive and negative emotions is related to distinct patterns of heart rate variability and situational empathy. *PLoS One* 15:e0244427. doi: 10.1371/journal.pone.0244427

Lamm, C., Nusbaum, H. C., Meltzoff, A. N., and Decety, J. (2007). What are you feeling? Using functional magnetic resonance imaging to assess the modulation of sensory and affective responses during empathy for pain. *PLoS One* 2:e1292. doi: 10.1371/journal.pone.0001292

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 261–273. doi: 10.1111/j.1469-8986.1993.tb03352.x

Laroche, J., Berardi, A. M., and Brangier, E. (2014). Embodiment of intersubjective time: relational dynamics as attractors in the temporal coordination of interpersonal behaviors and experiences. *Front. Psychol.* 5:1180. doi: 10.3389/fpsyg.2014.01180

Lebert, A., Chaby, L., Garnot, C., and Vergilino-Perez, D. (2020). The impact of emotional videos and emotional static faces on postural control through a personality trait approach. *Exp. Brain Res.* 238, 2877–2886. doi: 10.1007/s00221-020-05941-5

Lelard, T., Godefroy, O., Ahmaidi, S., Krystkowiak, P., and Mouras, H. (2017). Mental simulation of painful situations has an impact on posture and psychophysiological parameters. *Front. Psychol.* 8:2012. doi: 10.3389/fpsyg.2017.02012

Lelard, T., Krystkowiak, P., Montalan, B., Longin, E., Bucchioni, G., Ahmaidi, S., et al. (2014). Influence of postural threat on postural responses to aversive visual stimuli. *Behav. Brain Res.* 266, 137–145. doi: 10.1016/j.bbr.2014.02.051

Lelard, T., Stins, J., and Mouras, H. (2019). Postural responses to emotional visual stimuli. *Neurophysiol. Clin.* 49, 109–114. doi: 10.1016/j.neucli.2019.01.005

Martínez-Pernía, D., Cea, I., and Kaltwasser, A. (2021) in *Recovering the phenomenological and intersubjective nature of mindfulness through the enactive approach. Relational mindfulness: Fundamental and applications*. eds. R. Arístegui, J. García-Campayo and P. Barriga (Springer, Switzerland), 65–89.

Martínez-Pernía, D., Cea, I., Troncoso, A., Blanco, K., Calderón Vergara, J., Baquedano, C., et al. (2023). 'I am feeling tension in my whole body': an experimental phenomenological study of empathy for pain. *Front. Psychol.* 13:999227. doi: 10.3389/fpsyg.2022.999227

Martínez-Pernía, D., Rivera-Rei, Á., Forno, G., Troncoso, A., Aravena, O., Vergara, M., et al. (2020). A study based on the embodied emotion approach: the recognition of whole body social emotions and the postural control in Alzheimer's disease dementia, Parkinson's disease and healthy control. *Alzheimers & Dementia*. 16:e047687. doi: 10.13140/RG.2.2.17940.12165

Mouras, H., and Lelard, T. (2021). Approach-avoidance behavior in the empathy for pain model as measured by posturography. *Brain Sci.* 11:1426. doi: 10.3390/brainsci11111426

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Neal, D. T., and Chartrand, T. L. (2011). Embodied emotion perception. *Soc. Psychol. Personal. Sci.* 2, 673–678. doi: 10.1177/1948550611406138

Newen, A., Gallagher, S., and De Bruin, L. (2018). 4E cognition. *Oxford Handbook Cognition* 4E. in Oxford Library of Psychology. Oxford Academic. 2–16. doi: 10.1093/oxfordhb/9780198735410.013.1

Oberman, L. M., Winkielman, P., and Ramachandran, V. S. (2007). Face to face: blocking facial mimicry can selectively impair recognition of emotional expressions. *Soc. Neurosci.* 2, 167–178. doi: 10.1080/17470910701391943

Olivares, F. A., Vargas, E., Fuentes, C., Martínez-Pernía, D., and Canales-Johnson, A. (2015). Neurophenomenology revisited: second-person methods for the study of humanconsciousness. *Front. Psychol.* 6:673. doi: 10.3389/fpsyg.2015.00673

Peled-Avron, L., Goldstein, P., Yellinek, S., Weissman-Fogel, I., and Shamay-Tsoory, S. G. (2018). Empathy during consoling touch is modulated by mu-rhythm: an EEG study. *Neuropsychologia* 116, 68–74. doi: 10.1016/j.neuropsychologia.2017.04.026

Petitmengin, C. (2006). Describing one's subjective experience in the second person: an interview method for the science of consciousness. *Phenomenol. Cogn. Sci.* 5, 229–269. doi: 10.1007/s11097-006-9022-2

Petitmengin, C., Remillieux, A., and Valenzuela-Moguillansky, C. (2018). Discovering the structures of lived experience. *Phenomenol. Cogn. Sci.* 18, 691–730. doi: 10.1007/s11097-018-9597-4

Pizarro, D., Zepeda, A., and Martínez-Pernía, D. (2023). Estudio fenomenológico experimental de la empatía por dolor en adulto mayor sano. *Limite (Chile).* 18, Artículo, 24.

Reed, C. L., Moody, E. J., Mgrublian, K., Assaad, S., Schey, A., and McIntosh, D. (2020). Body matters in emotion: restricted body movement and posture affect expression and recognition of status-related emotions. *Front. Psychol.* 11:1961. doi: 10.3389/fpsyg.2020.01961

Riečanský, I., and Lamm, C. (2019). The role of sensorimotor processes in pain empathy. *Brain Topogr.* 32, 965–976. doi: 10.1007/s10548-019-00738-4

Spielberger, C. D., Gorsuch, R. C., and Lushene, R. E. (1970). *Manual for the state trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press

Tanaka, S. (2015). Intercorporeality as a theory of social cognition. *Theor. Psychol.* 25, 455–472. doi: 10.1177/0959354315583035

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Massachusetts: Harvard University Press.

Troncoso, A., Soto, V., Gomila, A., and Martínez-Pernía, D. (2023). Moving beyond the lab: investigating empathy through the empirical 5E approach. *Front. Psychol.* 14:1119469. doi: 10.3389/fpsyg.2023.1119469

Uithol, S., and Gallese, V. (2015). The role of the body in social cognition. *In WIREs Cogn. Sci.* 6, 453–460. doi: 10.1002/wcs.1357

Valenzuela-Moguillansky, C., and Vásquez-Rosati, A. (2019). An analysis procedure for the micro-phenomenological interview. *Constr. Found.* 14, 123–145. Available at: https://constructivist.info/14/2/123.

Varela, F. J. (1996). Neurophenomenology: a methodological remedy for the hard problem. *J. Conscious. Stud.* 3, 330–349.

Varela, F. J., Rosch, E., and Thompson, E. (1991). *The embodied mind*. Massachusetts: MIT Press.

Vergara, M., Cea, I., Calderón, J., Troncoso, A., and Martínez-Pernía, D. (2022). "An experimental phenomenological approach to the study of inner speech in empathy: bodily sensations, emotions, and felt knowledge as the experiential context of inner spoken voices" in *New perspectives on inner speech*. ed. P. Fossa (Cham: Springer), 65–80.

Weissman, D. G., and Mendes, W. B. (2021). Correlation of sympathetic and parasympathetic nervous system activity during rest and acute stress tasks. *Int. J. Psychophysiol.* 162, 60–68. doi: 10.1016/j.ijpsycho.2021.01.015

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 27 March 2024
DOI 10.3389/fpsyg.2024.1322781

OPEN ACCESS

EDITED BY
Chiara Lucifora,
University of Bologna, Italy

REVIEWED BY
Francesca Ciardo,
University of Milano-Bicocca, Italy
Kostas Karpouzis,
Panteion University, Greece

*CORRESPONDENCE
Rose E. Guingrich
✉ rose.guingrich@princeton.edu

RECEIVED 16 October 2023
ACCEPTED 13 March 2024
PUBLISHED 27 March 2024

CITATION
Guingrich RE and Graziano MSA (2024) Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction.
Front. Psychol. 15:1322781.
doi: 10.3389/fpsyg.2024.1322781

Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction

Rose E. Guingrich[1,2]* and Michael S. A. Graziano[1,3]

[1]Department of Psychology, Princeton University, Princeton, NJ, United States, [2]Princeton School of Public and International Affairs, Princeton University, Princeton, NJ, United States, [3]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, United States

The question of whether artificial intelligence (AI) can be considered conscious and therefore should be evaluated through a moral lens has surfaced in recent years. In this paper, we argue that whether AI is conscious is less of a concern than the fact that AI can be considered conscious by users during human-AI interaction, because this ascription of consciousness can lead to carry-over effects on human-human interaction. When AI is viewed as conscious like a human, then how people treat AI appears to carry over into how they treat other people due to activating schemas that are congruent to those activated during interactions with humans. In light of this potential, we might consider regulating how we treat AI, or how we build AI to evoke certain kinds of treatment from users, but not because AI is inherently sentient. This argument focuses on humanlike, social actor AI such as chatbots, digital voice assistants, and social robots. In the first part of the paper, we provide evidence for carry-over effects between perceptions of AI consciousness and behavior toward humans through literature on human-computer interaction, human-AI interaction, and the psychology of artificial agents. In the second part of the paper, we detail how the mechanism of schema activation can allow us to test consciousness perception as a driver of carry-over effects between human-AI interaction and human-human interaction. In essence, perceiving AI as conscious like a human, thereby activating congruent mind schemas during interaction, is a driver for behaviors and perceptions of AI that can carry over into how we treat humans. Therefore, the fact that people can ascribe humanlike consciousness to AI is worth considering, and moral protection for AI is also worth considering, regardless of AI's inherent conscious or moral status.

KEYWORDS

artificial intelligence, human-AI interaction, theory of mind, consciousness, schemas, chatbots

Introduction

Consciousness is considered the subjective experience that people feel in association with events, such as sensory events, memories, and emotions (Nagel, 1974; Harley, 2021). Many people study consciousness, and there are just as many competing theories about what it is and how it is generated in the human brain (e.g., Chalmers, 1996; Baars, 1997; Tononi, 2007;

Graziano, 2013; Doerig et al., 2020). Recently, people have speculated that artificial intelligence can also have consciousness (e.g., O'Regan, 2012; Yampolskiy, 2018; Chalmers, 2023). Whether that is possible, and how, is still debated (e.g., Koch, 2019). However, it is undeniable that children and adults attribute consciousness to AI through Theory of Mind attributions (Kahn et al., 2012; Broadbent et al., 2013; Eyssel and Pfundmair, 2015; Martini et al., 2016; Tanibe et al., 2017; Świderska and Küster, 2018; Heyselaar and Bosse, 2020; Küster and Świderska, 2020; Taylor et al., 2020). Some researchers have argued that consciousness is fundamentally an attribution, a construct of social cognitive machinery, and that we attribute it to other people and to ourselves (Frith, 2002; Graziano, 2013; Prinz, 2017). As such, regardless of whether AI is conscious, attributing consciousness to AI matters in the same way attributing it to other humans does.

Premack and Woodruff (1978) coined the term Theory of Mind (ToM), which is the ability to attribute mind states to oneself and others more expansive. For example, one heavily studied aspect of ToM is the ability to recognize false beliefs in others (Wimmer and Perner, 1983). This cognitive capability has historically distinguished humans from many other species, yet Rabinowitz et al. (2018) claimed that artificial intelligence passed the false belief test. ToM may extend beyond attributing beliefs to attributing other aspects of mind such as emotions and intentionality. According to some, ToM can be divided into two distinct processes: attributing agency, or the ability to decide and act autonomously, and attributing experience, or the ability to have subjective states (Gray et al., 2007; Knobe and Prinz, 2007). Attributing consciousness to AI is therefore probably not one, single process, but instead should be broken down into experience and agency, with each part analyzed separately (Ward et al., 2013; Küster et al., 2020).

It has been suggested that attributing experience, rather than agency, plays a larger role in the perception of consciousness in AI (Knobe and Prinz, 2007). This distinction may present some difficulties for accurately measuring whether people view AI as conscious. People are generally more willing to assign agency rather than experience to a variety of targets, including robots (Gray and Wegner, 2012; Jacobs et al., 2021). This may be due in part to it being easier to determine whether an agent can make decisions or act on its own (agency) than whether an agent can feel pain or pleasure (experience). Adding further complexity, not all people ascribe agency and experience to AI in the same manner. For example, psychopathology and personality traits such as emotional stability and extraversion correlate with whether someone ascribes agency or experience to robots: emotional stability positively correlates with ascribing agency to robots, and extraversion positively correlates with attributing experience to robots (Tharp et al., 2016). Other individual differences such as people's formal education may also relate to whether someone attributes agency characteristics like intentionality to a humanoid robot (Roselli et al., 2023). Given these findings, it may be useful to operationalize ToM as a complex, overarching collection of interrelated processes, each of which plays a different role in how people attribute consciousness to machines.

The attribution of consciousness to AI is particularly relevant to social actor AI. These humanlike agents are social embodiments of intelligent algorithms that people can talk to and even engage with physically. Social actor AI includes chatbots, digital voice assistants, and social robots. Social actor AI's humanlike characteristics, from how the AI is embodied—like its bodily form, voice, and even linguistic style—to its ability to process social information, are unique within the category of artificial, non-human agents. Social actor AI is arguably more akin to humans than are other machines and objects. As such, how people behave toward social actor AI agents might be more likely to impact how they behave toward another human, despite the fact that these AI agents are not themselves living beings. Velez et al. (2019) posited that "an increasingly important question is how these social responses to agents will influence people's subsequent interactions with humans." Moreover, social actor AI is evolving rapidly. As Etzrodt et al. (2022) described it, "We are witnessing a profound change, in which communication *through* technologies is extended by communication *with* technologies." Instead of using social media as a medium through which you can interact with other people, users can, for example, download an app through which they can interact with a non-human being. Companion chatbots like Replika, Anima, or Kiku have millions of people using their apps. Millions more have digital voice assistants such as Siri and Alexa operating on their smartphones and in their homes. People form relationships with these agents and can come to view them as members of the family, friends, and even lovers (Croes and Antheunis, 2020; Garg and Sengupta, 2020; Brandtzæg et al., 2022; Xie and Pentina, 2022; Guingrich and Graziano, 2023; Loh and Loh, 2023). AI agents will almost certainly become both more ubiquitous and humanlike. As new generations grow up with these technologies on their mobile devices and in their homes, the consequences of humanlike AI will likely become more pronounced over time.

In this paper, we will not consider what, exactly, consciousness is, what causes it, or whether non-human machines can have it. Instead, the goal here is to discuss how people perceive consciousness in social actor AI, to explore the possible profound social implications, and to suggest potential research questions and regulatory considerations for others to pursue within this scope of research.

Part 1: evidence for carry-over effects between human-AI interaction and human-human interaction

Carry-over effects between AI's tangible and intangible characteristics

When people interact with AI, tangible characteristics of the agent such as appearance or embodiment, behavior, communication style, gender, and voice can affect how people perceive intangible characteristics such as mind and consciousness, emotional capability, trustworthiness, and moral status (Powers and Kiesler, 2006; Gray and Wegner, 2012; Broadbent et al., 2013; Eyssel and Pfundmair, 2015; Seeger and Heinzl, 2018; Lee et al., 2019; Küster et al., 2020; Dubosc et al., 2021; Rhim et al., 2022). The critical tangible-intangible relationship examined here is the one between an agent's humanlike embodiment and consciousness ascription (Krach et al., 2008; Broadbent et al., 2013; Ferrari et al., 2016; Abubshait and Wiese, 2017; Stein et al., 2020).

Generally, the more tangibly humanlike that people perceive an AI agent to be, the more likely people are to ascribe mind to the agent (e.g., Broadbent et al., 2013). At least one study suggests that mind ascription does not increase with human likeness until a particular threshold of human likeness is reached; once an agent's appearance

reaches the middle of the machine-to-human spectrum and the AI agent's appearance includes actual human features such as eyes and a nose, then mind ascription begins to increase with human likeness (Martini et al., 2016).

People are not always aware that they attribute mind to an AI agent during interaction. In other words, the construct of mind or consciousness activated in people during these interactions may be implicit, making it more difficult to measure. Banks (2019) conducted an online survey to compare participants' implicit and explicit ascriptions of mind to an agent. Participants ($N = 469$) were recruited from social media and university research pools and were randomly assigned to one of four agents. Three of the agents were social AIs that varied in their human likeness and mind capacity, and one was a human control, all named "Ray." Banks tested implicit ascription of mind using five classic ToM tests that measure whether participants ascribe mind to an agent including the white lie scenario and the Sally-Anne test. Explicit measures of mind were measured by two questions: do you think Ray has a mind, and how confident are you in your response? For the implicit tests' open-ended responses, trained, independent raters coded the data for mentalistic explanations of behavior. The results showed that while people implicitly ascribed ToM to humanlike AI, this implicit ascription did not correlate with explicit mind ascriptions.

Mind ascription appears to be automatically induced by AI's tangible human likeness, even when subjects are prompted to believe the opposite. Stein et al. (2020) compared mind ascriptions in a 2×2 between-subjects design of embodiment and mind capability for 134 German-speaking participants recruited from social media and mailing lists. Stimuli included vignettes and videos of either a text-based chatbot interface (Cleverbot) or a humanoid robot (with a 3-D rendered face of a woman) that was described as built on a simple or complex algorithm. The complex algorithm description included humanlike mind traits such as empathy, emotions, and understanding of the user. The researchers found a multivariate main effect of embodiment, such that people ascribed more mind capabilities to the humanoid robot than the text-based chatbot, regardless of whether it was based on a simple or complex algorithm. These researchers reported that "a digital agent with human-like visual features was indeed attributed with a more human-like mind—regardless of the cover story that was given regarding its actual mental prowess."

In sum, evidence suggests that an AI agent's observable or tangible characteristics, specifically its humanlike appearance, leads automatically to ascribing intangible characteristics, including consciousness, to the AI agent. As such, slight adjustments to AI's tangible characteristics can impact whether people perceive the artificial agent as conscious.

Carry-over effects between perceiving mind in AI and human-AI interaction

In some cases, ascribing a mind to AI is linked with viewing the agent as likable and trustworthy (Young and Monroe, 2019), which can impact whether people engage in helping behaviors. Srinivasan and Takayama (2016) found that when people perceived a robot as having an agentic mind, such that the robot was acting of its own accord rather than being controlled by a human, they came to its aid 50% more quickly. Study 1 was a mixed experiment design conducted online ($N = 354$, recruited from Amazon Mechanical Turk) in which participants each watched eight videos of robots requesting help using various politeness strategies, and study 2 was a behavioral lab study ($N = 48$, recruited via university participant pools and postings in local areas) with three conditions that were based on study 1's results. In study 2, participants watched a movie with a robot in the room (Willow Garage's Personal Robot 2). During the movie, the robot brought food to the participant and mentioned that the room looked like it needed to be cleaned, offered to do so, and requested aid from the participant. While the majority of participants helped the robot, those participants who rated the robot as more agentic came to its aid more quickly.

Depending on the paradigm, ascribing mind to AI can affect ease of interaction by augmenting or inhibiting the dyadic flow. Interacting with a humanlike artificial agent spurs the automatic use of human social scripts (Nass and Moon, 2000; Nass and Brave, 2005) and other social processes (von der Pütten et al., 2009), which can facilitate human-AI interaction (Sproull et al., 1996; Rickenberg and Reeves, 2000; Krämer et al., 2003a,b; Duffy, 2008; Krämer et al., 2009; Vogeley and Bente, 2010; Kupferberg et al., 2011). Facilitation of interaction and likability are however dependent on individual differences such as familiarity with the AI (Wang et al., 2021), need for social inclusion or interaction (Lee et al., 2006; Eyssel and Pfundmair, 2015), and other individual differences (Lee, 2010).

At a certain point, interaction facilitation no longer increases with human likeness across both tangible and intangible domains. The benefits of human likeness decrease dramatically when human likeness suddenly becomes creepy, according to the Uncanny Valley Hypothesis coined by Mori (1970). When an AI agent's appearance approaches the tipping point of "not enough machine, not enough human," the AI has entered the dip of the uncanny valley. At this point, an artificial agent's human likeness becomes disturbing, thereby causing anxiety or discomfort in users. The discomfort arising from the uncanny valley effect is generally distinct from dislike yet can have similar negative effects on the flow of interaction (Quadflieg et al., 2016).

The uncanny valley theory of human-AI interaction more recently acquired a qualifier: the uncanny valley of *mind* (Stein and Ohler, 2017; Appel et al., 2020). No longer just concerned with general human likeness, the uncanny valley effect can occur when AI's mind capabilities get too close to that of a human mind. It is uncertain whether negative uncanny valley effects of mind are stable, however, given the contradictions within this more recent scope of research. In Stein et al.'s study, they also found that the AI with low mind capacity, based on a simple algorithm rather than an advanced one, caused more discomfort when the AI was embodied rather than solely text-based. In another study, the researchers found that the more people perceived AI *or* humans to have a typically human mind, the less eerie feelings they experienced (Quadflieg et al., 2016). Due to inconsistent stimuli across studies, it is possible that slight variations in facial features or voice of the AI agent drove these dissimilar effects. In these cases, it may be useful to control for appearance when attempting to parse out the impacts of the uncanny valley of mind on how people interact with AI agents.

Via a series of three studies, Gray and Wegner (2012) made the claim that experiential aspects of mind, and not those of agentic mind, drive uncanny valley effects. In one of the studies, participants, recruited from subway stations and dining halls ($N = 45$), were given

vignettes of a supercomputer that was described as having only experience capabilities, having only agency, or simply mechanical. They then rated their feelings (uneasy, unnerved, and creeped out) and perceptions of the supercomputer's agency and experience. The experiential supercomputer elicited significantly higher uncanny valley feelings than agents in the other two conditions. Apparently, an intelligent computer that is seen as having emotion is creepier than one that can make autonomous decisions. The distinction between uncanny valley effects of experience and agency may be caused by feelings of threat: AI agents that are capable of humanlike emotion threaten that which makes mankind special (Stein and Ohler, 2017). If threat drove discomfort in Gray and Wegner's participants, then familiarity with the agent might mitigate perceptions of threat to the point at which the uncanny valley switches into the "happy valley." According to that hypothesis, after long-term, comfortable, and safe exposure to a humanlike AI agent, people might find the agent's human likeness to increase its likability, which might facilitate human-AI interaction (Cheetham et al., 2014).

The uncanny valley effect with respect to AI is therefore more complicated and difficult to study than it may at first appear. Familiarity with AI over time, combined with the increasing ubiquity of social actor AI, may eliminate uncanny valley effects altogether. Uncanny valley effects differ across studies, and are affected by multiple factors, including expectation violation (Spence et al., 2014; Edwards et al., 2019; Lew and Walther, 2022), individual differences (MacDorman and Entezari, 2015), and methodological differences such as stimuli and framing. Further, the way the uncanny valley graph rises to a peak has been contested. For example, researchers have debated exactly where that peak lies on the machine-to-human scale (Cheetham et al., 2014; Pütten and Krämer, 2014; Stein et al., 2020). However, what we do know is that perceiving mind in AI affects people's emotional state and how they interact with AI, making the intangible characteristic of mind one of the mechanisms that impacts human-AI interaction.

Carry-over effects between human-AI interaction and human-human interaction

Most studies on human-AI interactions, such as those reviewed above, focus on what could be called one-step effects like the uncanny valley effect, trust, and likability. Such studies are concerned with how characteristics of AI impact how people interact with the agent. Arguably a more important question is the two-step effect of how human-AI interactions might impact subsequent human-human interactions. Though findings on these two-step effects are limited and sometimes indirect, the data do suggest that such effects are present. The impact of AI is not confined to the interaction between a user and an AI agent, but rather carries over into subsequent interactions between people.

Social Cognitive Theory, anthropomorphism, and ToM literature provide theoretical foundations for why interactions with social actor AI could prompt carry-over effects on human-human interaction. Due to the social nature of these agents, AI can act as a model for social behavior that users may learn from (Bandura, 1965, 1977). According to Waytz et al. (2010), when someone anthropomorphizes or ascribes mind to an artificial agent, that agent then "serves as a source of social influence on the self." In other words, "being watched by others matters, perhaps especially when others have a mind like one's own." Social actor AI is an anthropomorphized target; therefore, it can serve as a role model or operate as an ingroup member that has some involvement in setting social norms, as seen with the persuasive chatbot that convinced people to donate less to charity (Zhou et al., 2022), the chatbot that persuaded users to get vaccinated for COVID-19 or participate in social distancing (Kim and Ryoo, 2022), and the humanlike avatar that elicited more socially desirable responses from participants than a mere text-based chatbot did (Krämer et al., 2003a). Social actor AI can persuade people in these ways, regardless of whether people trust it or perceive it as credible (Lee and Liang, 2016, 2019). In some paradigms, chatbot influence mimics that of people: chatbots can implement foot-in-the-door techniques to influence people's emotions and bidding behavior in gambling (Teubner et al., 2015) and can alter consumers' attitudes and purchasing behavior (Han, 2021; Poushneh, 2021).

Another explanation for why AI can socially influence people may be that the user views the agent as being controlled by another human. Some research suggests that perceiving a human in the loop during interactions with AI results in stronger social influence and more social behavior (Appel et al., 2012; Fox et al., 2014). This idea, however, has since been contested (Krämer et al., 2015). Indeed, early research on human-computer interaction found that when people perceived a computer as a social agent, they did not simply view it as a product of human creation, nor did they imagine that they were interacting with the human engineer who created the machine (Nass et al., 1994; Sundar and Nass, 2000). Nass and colleagues designed a series of paradigms in which participants were tutored, via audio emitting from computer terminals, by computers or human programmers that subsequently evaluated participants' performance. To account for the novelty of computers at this time, earlier studies were conducted with experienced computer users. They found significant differences between computer and human tutor conditions, such that people viewed computers as not just entities controlled by human programmers, but entities to which the ideas of "self" and "other" and social agency applied. Nass and colleagues laid the groundwork for evaluating social consequences of interacting with intelligent machines, as their experiments provided initial evidence that people treated the machines themselves as social actors. As such, it may be the case that social influence is strengthened when people think a human is involved, yet social influence still exists when the AI agent is perceived as acting on its own accord.

Communication researchers have found that the way people communicate with AI is linked to how they communicate with other humans thereafter, such that people are then more likely to speak to another human in the same way in which they habitually speak to an artificial agent. For example, talking with the companion chatbot Replika caused users' linguistic styles to converge with the style of their chatbot over time (Wilkenfeld et al., 2022). The way children speak with social actor AI such as the home assistant, Alexa, can carry over into how children speak to their parents and others (Hiniker et al., 2021). Garg and Sengupta (2020) tracked and interviewed 18 families over an average of 58 weeks who used a digital voice assistant in their homes and analyzed raw audio interactions with their assistant. These researchers found that "when children give commands at

a high volume, there is an aggressive tone, which often unintentionally seeps into children's conversations with friends and family." A parent in the study commented that, "If I do not listen to what my son is saying, he will just start shouting in an aggressive tone. He thinks, as Google responds to such a tone, I would too." While home assistants can negatively impact communication, they can also foster communication within families and alter how communication breakdowns are repaired (Beneteau et al., 2019, 2020). Parents have concerns about their children interacting with social actor AI, but they also see AI's potential to support children by "attuning to others, cultivating curiosity, reinforcing politeness, and developing emotional awareness" (Fu et al., 2022). According to the observational learning concept in Social Cognitive Theory (Bandura, 1965), assistants might provide models for prosocial behavior that children could learn from (such as being polite, patient, and helpful) regardless of whether the assistant provides positive reinforcement when children act in these prosocial ways. The studies mentioned above show how both children's positive and negative modes of communication can be reinforced via interactions with home assistants.

Not only can social actor AI affect the way that people communicate with each other within their relationships, but also it has the potential to impact relationships with other people due to attachment to the agent. Through in-depth interviews of existing Replika users ($N = 14$, ages 18–60), Xie and Pentina (2022) suggested that AI companions might replace important social roles such as family, friends, and romantic partners through unhealthy attachment and addiction. An analysis of families' use of Google Home revealed that children, specifically those between the age of 5–7, believed the device to have feelings, thoughts, and intentions and developed an emotional attachment to it (Garg and Sengupta, 2020). These children viewed Google Home as if it had a mind through ascribing characteristics of agency and experience to it.

The psychosocial benefits of interactions with social actor AI may either contribute to positive relational skill-building if AI is used as a tool, or they may lead to human relationship replacement if these benefits are comparatively too difficult to get from relationships with real people. Research suggests that people self-disclose more when interacting with a computer versus with a real person, in part due to people having lower fear of being judged, thereby prompting more honest answers (Lucas et al., 2014). This effect is found even though benefits of emotional self-disclosure are equal whether people are interacting with chatbots or human partners (Ho et al., 2018). Further, compared to interacting with other people, those interacting with artificial agents experience fewer negative emotions and lower desire for revenge or retaliation (Kim et al., 2014). Surveys of users of the companion chatbot, Replika, suggest that users find solace in human-chatbot relationships. Specifically, those who have experienced trauma in their human relationships, for example, indicate that Replika provides a safe, consistent space for positive social interaction that can benefit their social health (Ta et al., 2020; Guingrich and Graziano, 2023). The question is whether the benefits of human-AI interaction presented here may lead to people choosing AI companions over human ones.

In part 1, we have reviewed evidence that human-AI interaction, when moderated by perceiving the agent as having a humanlike mind or consciousness, has carry-over effects on human-human interaction. In part 2, we address the mechanism of this moderator through congruent schema activation. We further pose two theoretical types of carry-over effects that may occur via congruent schema activation: relief and practice.

Part 2: mechanisms and types of carry-over effects: schemas and relief or practice

Schema congruence and categorization

What is the mechanism by which people's attributions of consciousness to AI lead to carry-over effects on interactions with other humans? One possibility is the well-known mechanism of activating similar schemas of mind when interacting with different agents. We propose that ascribing mind or consciousness to AI through automatic, congruent schema activation is the driving mechanism for carry-over effects between human-AI interaction and human-human interaction.

Schemas are mental models with identifiable properties that are activated when engaging with an agent or idea and are useful ways of organizing information that help inform how to conceptualize and interact with new stimuli (Ortony and Anderson, 1977; McVee et al., 2005; Pankin, 2013). For example, the schema you have for your own consciousness informs how you understand the consciousness of others. You assume, because your experience of consciousness contains X and Y characteristics, that another person's consciousness also contains X and Y characteristics, and this facilitates understanding and subsequent social interaction between you and the other person (Graziano, 2013).

Researchers have analyzed the consequences of failing to fully activate all properties of mind schemas between similar agents. For example, the act of dehumanization reflects a disconnect between how you view your mind and that of other people. Instead of activating the consciousness schema with X and Y characteristics during interaction with another human, you may activate only the X characteristic of the schema. Dehumanization is linked to social consequences such as ostracism and exclusion, which can harm social interaction (Bastian and Haslam, 2010; Haslam and Loughnan, 2014).

We can apply the idea of schema congruence to interactions with social actor AI while also taking into consideration the level of advancement of the AI in question. Despite AI being more advanced than other technology like personal mobile devices or cars in terms of human likeness and mind ascription, some research suggests that social actor AI still falls short of the types of mind schemas that are activated when people interact with each other. However, humanlike AI is developing at a rapid rate. As it does, the schematic differences between AI agents and humans will likely blur more than they already have. To better understand the consequences of current social actor AI, it may be prudent to observe the impacts of human-AI interaction through ingroup-outgroup or dehumanization processes, both of which are useful psychological lenses for group categorization. We propose that psychological tests of mind schema activation will be especially useful for more advanced, future AI that is more clearly different from possessions like cars and phones but similar to humans in terms of mind characteristics.

Schematic incongruence yields uncanny valley effects

Categorization literature attempts to delineate whether people treat social actor AI as non-human, human, or other. The data are mixed, but some of the results may stem from earlier AI that is not as capable. Now that AI is becoming sophisticated enough that people can more easily attribute mind to it, the categories may change. In this literature, social AI is usually classified by study participants as somewhere on the spectrum between machine and human, or it is classified as belonging to its own, separate category (Severson and Carlson, 2010). That separate category is often described as not quite machine, not quite human, with advanced communication skills and other social capabilities, and has been labeled with mixed-category words like humanlike, humanoid, and personified things (Etzrodt and Engesser, 2021).

Some researchers claim that the uncanny valley effect is driven by categorization issues. In that hypothesis, humanlike AI is creepy because it does not fit into categories for machine or human but exists in a space for which people do not have a natural, defined category (Burleigh et al., 2013; Kätsyri et al., 2015; Kawabe et al., 2017). Others claim that category uncertainty is not the driver of the uncanny valley effect, but, rather, inconsistency is (MacDorman and Chattopadhyay, 2016). In that hypothesis, because of the inconsistencies between AI and the defining features of known categories, people treat humanoid AI agents as though they do not fit into a natural, existing category (Gong and Nass, 2007; Kahn et al., 2011). Because social actor AI defies boundaries, it may trigger outgroup processing effects such as dehumanization that contribute to negative affect. The cognitive load associated with category uncertainty, more generally, may also trigger negative emotions that are associated with the uncanny valley effect.

Social norms likely play a role in explicit categorization of social AI (Hoyt et al., 2003). People may be adhering to a perceived social norm when they categorize social AI as machinelike rather than humanlike. It is possible that people explicitly place AI into a separate category from people, while the implicit schemas activated during interaction contradict this separation. The uneasy feeling from the uncanny valley effect may be a product of people switching between ascribing congruent mind schemas to the agent in one moment and incongruent ones in the next.

Schematic congruence yields carry-over effects on human-human interaction

As humanlike AI approaches the human end of the machine-to-human categorization spectrum, it also advances toward a position in which people can more easily ascribe a conscious mind to it, thereby activating congruent mind schemas during interactions with it. Activating congruent schemas impacts how people judge the agent and its actions. For example, the belief that you share the same phenomenological experience with a robot changes the way you view its level of intent or agency (Marchesi et al., 2022). Activation of mind-similarity may resemble simulation theory (Harris, 1992; Röska-Hardy, 2008). In that hypothesis, the observer does not merely believe the artificial agent has a mind but simulates that mind through the neural machinery of the person's own mind. Simulation allows the agent to seem more familiar, which facilitates interaction.

Some researchers have used schemas as a lens to explain why people interact differently with computer partners vs. human ones (Hayashi and Miwa, 2009; Merritt, 2012; Velez et al., 2019). In this type of research, participants play a game online and are told that their teammate is either a human or a computer, but, unbeknownst to the participants, they all interact with the same confederate-controlled player. This method allows researchers to observe how schemas drive perceptions and behavior, given that the prime is the only difference. According to Fox et al. (2014), when people believed themselves to be interacting with a human agent, they were more likely to be socially influenced. Velez et al. (2019) took this paradigm one step further and observed that activating schemas of a human mind during an initial interaction with an agent resulted in carry-over effects on subsequent interactions with a human agent. These researchers employed a 2×2 between-subjects design in which participants played a video game with a computer agent or human-backed avatar. They then were presented with the option to engage prosocially through a prisoner's dilemma money exchange with a stranger thereafter. When participants ($N = 184$) thought they were interacting with a human and that player acted pro-socially, they behaved more pro-socially toward the stranger. However, when participants believed they were interacting with a computer-controlled agent and it behaved pro-socially toward them, they had lower expectations of reciprocity and donated less game credits to the human stranger with whom they interacted subsequently. In the interpretation of Velez et al., the automatic anthropomorphism of the computer-backed agent was a mindless process (Kim and Sundar, 2012) and therefore not compatible with the cognitive-load-requiring social processes thereafter (Velez et al., 2019).

One of the theories that arose from research on schema activation in gaming is the Cooperation Attribution Framework (Merritt, 2012). According to Merritt, the reason people behave differently when game playing with a human vs. an artificial partner is that they generate different initial expectations about the teammate. These expectations activate stereotypes congruent with the teammates' identity, and confirmations of those stereotypes are given more attention during game play, causing a divergence in measured outcomes. According to Merritt, "the differences observed are broadly the result of being unable to imagine that an AI teammate could have certain attributes (e.g., emotional dispositions). …the 'inability to imagine' impacts decisions and judgments that seem quite unrelated." The computer-backed agents used in this research may evoke a schema incompatible with humanness—one that aligns with the schema of a pre-programmed player without agency—whereas more modern, advanced AI might evoke a different, more congruent schema in human game players.

Other studies examined schema congruence by seeing how people interact with and perceive an AI agent if its appearance and behavior do not fit into the same humanlike category. Expectation violation and schema incongruence appear to impact social responses to AI agents. In two studies, Ciardo et al. (2021, 2022) manipulated whether an AI agent looked humanlike and made errors in humanlike (vs. mechanical) ways. They then observed whether people attributed intentionality to the agent or were socially inclusive with it.

Coordination with the AI agent during the task and social inclusion with the AI agent after the task were impacted by humanlike errors during the task only if the agent's appearance was also humanlike. This variation in response toward the AI may have to do with ease of categorization: if an agent looks humanlike and acts humanlike, the schemas activated during interaction are stable, which facilitates social response to the agent. On the other hand, if an agent looks humanlike but does not act humanlike, schemas may be switching and people may incur cognitive load and feel uncertain about how to respond to the agent's errors. In their other study, these researchers found that when a humanlike AI agent's mistakes were also humanlike, people attributed more intentionality to it than when a humanlike AI agent's mistakes were mechanical.

To understand why people might unconsciously or consciously view social actor AI as having humanlike consciousness, it is useful to understand individual differences that contribute to automatic anthropomorphism (Waytz et al., 2010) and therefore congruent schema activation. Children who have invisible imaginary friends are more likely to anthropomorphize technology, and this is mediated by what the researchers call the "imaginative process of simulating and projecting internal states" through role-play (Severson and Woodard, 2018). As social AI agents become more ubiquitous, it is likely that mind-ascription anthropomorphism will occur more readily; for instance, intensity of interaction with the chatbot Replika mediates anthropomorphism (Pentina et al., 2023). Currently, AI is not humanlike enough to be indistinguishable from real humans. People are still able to identify real from artificial at a level better than chance, but this is changing. What might happen once AI becomes even more humanlike to the point of being indistinguishable from real humans? At that point, the people who have yet to generate a congruent consciousness schema for social actor AI may do so. Others may respond by becoming more sensitive to subtle, distinguishing cues and by creating more distinct categories for humans and AI agents. At some point in the development of AI, perhaps even in the near future, the distinction between AI behavior and real human behavior may disappear entirely, and it may become impossible for people to accurately separate these categories no matter how sensitive they are to the available cues.

Possible types of carry-over effects: relief or practice

What, exactly, is the carry-over effect between human-AI interaction and human-human interaction? We will examine two types of carry-over effects that do not necessarily reflect all potential outcomes but that provide a useful comparison by way of their consequences: relief and practice. In the case of relief, doing X behavior with AI will cause you to do less of X behavior with humans subsequently. In the case of practice, doing X behavior with AI will cause you to do more of X behavior with humans subsequently. The preponderance of the evidence so far suggests that practice is more likely to be observed, and its consequences outweigh those of relief (Garg and Sengupta, 2020; Hiniker et al., 2021; Wilkenfeld et al., 2022).

The following scenarios illustrate theoretical examples of both effects. Consider an example of relief. You are angry, and you let out your emotions on a chatbot. Because the chatbot has advanced communication capabilities and can respond intelligently to your inputs, you feel a sense of relief from berating something that reacts to your anger. Over time, you rely on ranting to this chatbot to release your anger, and as a result, you are relieved of your negative emotions and are less likely to lash out at other people.

Now consider an example of practice. Suppose you are angry. You decide to talk to a companion chatbot and unleash your negative emotions on the chatbot, speaking to it rudely through name-calling and insults. The chatbot responds only positively or neutrally to your attacks, offering no negative backlash in return. This works for you, so you continue to lash out at the chatbot when angry. Since this chatbot is humanlike, you tend not to distinguish between this chatbot and other humans. Over time, you start to lash out at people as well, since you have not received negative feedback from lashing out at a humanlike agent. The risk threshold for relieving your anger at something that will socialize with you is decreased. You have effectively practiced negative behavior with a humanlike chatbot, which led to you engaging more in that type of negative behavior with humans. Practice can involve more than negative behaviors. Suppose you have a friendly, cooperative interaction with an AI, in which you feel safe enough to share your feelings. Having engaged in that practice, maybe you are more likely to engage in similar positive behavior to others in your life.

Both of these examples illustrate ways in which antisocial behavior toward humans can be reduced or increased by interactions with social actor AI. There are also situations in which prosocial behaviors can be reinforced. Which of the scenarios, relief or practice, are we more likely to observe? The answer to this question will inform the way society should respond to or regulate social actor AI.

Evidence against relief and evidence for practice effects

Researchers have proposed that people should take advantage of social actor AI's human likeness to use it as a cathartic object. Coined by Luria et al. (2020), the idea of a cathartic object is familiar: for example, a pillow can be used as a cathartic object by punching it in anger, thereby relieving oneself of the emotion. This is, colloquially, a socially acceptable behavior toward the target. Luria takes this one step further by suggesting that responsive, robotic agents that react to pain or other negative input can provide even more relief than an inanimate object, and that we should use them as cathartic objects. Luria claims that the reaction itself, which mirrors a humanlike pain response, provides greater relief than that of an object that does not react. One such "cathartic object" designed by Luria is a cushion that vibrates in reaction to being poked by a sharp tool. The more tools you put into the cushion, the more it vibrates until it shakes so violently that the tools fall out. You can repeat the process as much as desired.

The objects presented by Luria as potential agents of negative-emotion relief are simply moving, responsive objects at this stage. However, Luria proposes the use of more humanlike agents, such as social robots, as cathartic objects. In one such proposition, Luria suggests that people throw knives at a robotic, humanlike bust that responds to pain. In another example, Luria suggests a ceremonial interaction in which a child relieves negative emotions with a responsive robot that looks like a duck.

Luria's proposal rests on the assumption that releasing negative emotions on social robots will relieve the user of that emotion. Catharsis literature, however, challenges this assumption: research suggests that catharsis of aggression does not reduce subsequent aggression, but can in fact increase it, providing evidence for practice effects (Denzler and Förster, 2012; Konečni, 2016). Catharsis researchers posit that the catharsis of negative behavior and feelings requires subsequent training, learning, and self-development post-catharsis to lead to a reduction of the behavior. Therapy, for example, provides a mode through which patients can feel catharsis and then learn methods to reduce negative feelings or behaviors toward others. Even so, the catharsis or immediate relief alone does not promise a reduction of that behavior or feeling (Alexander and French, 1946; Dollard and Miller, 1950; Worchel, 1957) and can in many ways exacerbate negative feelings (Anderson and Bushman, 2002; Bushman, 2002). Other researchers found that writing down feelings of anger was less effective than writing to the person who made the participant angry, yet neither mode of catharsis alleviated anger responses (Zhan et al., 2021). These findings suggest that whether you were to write to a chatbot and tell it about your anger, or bully it, the behavior would only result in increased aggression toward other people.

Recent data on children and their interactions with home assistants such as Amazon's Alexa or Google Assistant suggest for plural data that negative interactions with AI, including using an aggressive, loud tone of voice with it, does not lead to a cathartic reduction in aggression toward others, but to the opposite, an increase in aggressive tone toward other people (Beneteau et al., 2019, 2020; Garg and Sengupta, 2020; Hiniker et al., 2021). This data suggests that catharsis does not work for children in their interactions with AI and may be cause for concern.

This concern is especially important given that children tend to perceive a humanlike mind in non-human objects in general, more so than adults. When asked to distinguish between living and non-living agents, including robots, children experience some difficulty. Even when children do not ascribe biological properties to robots, research suggests that children can still ascribe psychological properties, like agency and experience, to robots (Nigam and Klahr, 2000). There appears to be a historical trend of increasing mind ascription to technology in children over the years. This trend may reflect the increased human likeness and skills of technology, and therefore provide us a prediction for the future. In 1995, children at the age of five reported that robots and computers did not have brains like people (Scaife and Van Duuren, 1995), but in a research study in 2000, children ascribed emotion, cognitive abilities, and volition to robots, even though most did not consider the robot to be alive (Nigam and Klahr, 2000). In studies conducted in 2002 and 2003, children 3–4 years old tended not to ascribe experiential mind to robots but did ascribe agentic qualities such as the ability to think and remember (Mikropoulos et al., 2003). According to Severson and Woodard (2018), not unlike some theories of consciousness in which people perceive there to be a person inside their mind, "There are numerous anecdotes that young children think there's a little person inside the device" in home assistants like Alexa. Children with more exposure to and affinity with digital voice assistants have more pronounced psychological conceptions of technology, but it is unclear whether conceptions of technology and living things are blurred together (Festerling et al., 2022). Children do distinguish between technology and other living things through ascriptions of intelligence, however (Bernstein and Crowley, 2008). Goal-directed, autonomous behavior (a component of ToM) is one of the key mechanisms by which children distinguish an object as being alive (Opfer, 2002; Opfer and Siegler, 2004). Given that children appear to be ascribing mind to technology more than ever, this trend is likely to continue with AI advancement.

We are skeptical that socially mistreating AI can result in emotional relief, translating into better social behavior toward other people. Although the theory has been proposed, little if any evidence supports it. Encouraging people, and especially children, to berate or socially mistreat AI on the theory that it will help them become kinder toward people seems ill-advised to us. In contrast, the existing evidence suggests that human treatment of AI can sometimes result in a practice effect, which carries over to how people treat each other. Those practice effects could either result in social harm, if antisocial behavior is practiced, or social benefit, if pro-social behavior is practiced.

Discussion

The moral issue of perceiving consciousness in AI and suggested regulations

As stated at the beginning of this article, we do not take sides here on the question of whether AI is conscious. However, we argue that the fact that people often perceive it to be conscious is important and has social consequences. Mind perception is central to this process, and mind perception itself evokes moral thinking. Some researchers claim that "mind perception is the essence of morality" (Gray and Wegner, 2012). When people perceive mind in an agent, they may also view it as capable of having conscious experience and therefore perceive it as something worthy of moral care (Gray et al., 2007). Mind perception moderates whether someone judges an artificial agent's actions as moral or immoral (Shank et al., 2021). We suggest that when people perceive an agent to possess subjective experience, they perceive it to be conscious; when they perceive it to be conscious, they are more likely to perceive it as worthy of moral consideration. A conscious being is perceived as an entity that can act morally or immorally, and that can be treated morally or immorally.

We suggest it is worth at least considering whether social actor AI, as it becomes more humanlike, should be viewed as having the status of a moral patient or a protected being that should be treated with care. The crucial question may not be whether the artificial agent deserves moral protection, but rather whether we humans will harm ourselves socially and emotionally if we practice harming humanlike AI, and whether we will help ourselves if we practice pro-social behavior toward humanlike AI. We have before us the potential for cultural improvement or cultural harm as we continue to integrate social actor AI into our world. How can we ensure that we use AI for good? There are several options, some of which are unlikely and unenforceable, and one of which we view as being the optimal choice.

One option is to enforce how people treat AI, to reduce the risk of the public practicing antisocial behavior and to increase the

practice of prosocial behavior. Some have taken the stance that AI should be morally protected. According to philosophers such as Ryland (2021a,b), who characterizes relationships with robots in terms of friendship and hate, hate toward robots is morally wrong, and we should consider it even more so as robots become more humanlike. Others have claimed that we should give AI rights or protections, because AI inherently deserves them due to its moral-care status (Akst, 2023). Not only is this suggestion vague, but it is also pragmatically unlikely. Politically, it is overwhelmingly unlikely that any law would be passed in which a human being is supposed to be arrested, charged, or serve jail time for abusing a chatbot. The first politician to suggest it would end their career. Any political party to support it would lose the electorate. We can barely pass laws to protect transgender people; imagine the political and cultural backlash to any such legal protections for non-human machines. Regulating human treatment of AI is, in our opinion, a non-starter.

A second option is to regulate AI such that it discourages antisocial behavior and encourages prosocial behavior. We suggest this second option is much more feasible. For example, abusive treatment of AI by the user could be met with a lack of response (the old, "just ignore the bully and he'll go away, because he will not get the reaction he's looking for"). The industries backing digital voice assistants have already begun to integrate this approach into responses to bullying speech. In 2010, if a user told Siri, "You're a slut," it was programmed to respond with, "I'd blush if I could." Due to stakeholder feedback, the response has now been changed to a more socially healthy, "I will not respond to that" (UNESCO & EQUALS Skills Coalition et al., 2019; UNESCO, 2020). Currently, the largest industries backing AI, such as OpenAI with ChatGPT, are altering and restricting the types of inputs their social actor AI will respond to. This trend toward industry self-regulation of AI is encouraging. However, we are currently entirely dependent on the good intentions of industry leaders to control whether social actor AI encourages prosocial or antisocial behavior in users. Governing bodies have begun to make regulation attempts, but their proposals have received criticism: such documents try a "one-size-fits-all approach" that may result in further inequality. For example, the EU drafted an Artificial Intelligence Act (AIA) that proposes a ban on AI that causes psychological harm, but the potential pitfalls of this legislation appear to outweigh its impact on psychological well-being (Pałka, 2023).

Social actor AI is increasingly infiltrating every part of society, interacting with an increasing percentage of humanity, and therefore even if it only subtly shapes the psychological state and interpersonal behavior of each user, it could cause a massive shift of normative social behavior across the world. If there is to be government regulation of AI to reduce its risk and increase its benefit to humanity, we suggest that regulations aimed at its prosociality would make the biggest difference. One could imagine a Food and Drug Administration (FDA) style agency, informed by psychological experts, that studies how to build AI such that it reinforces prosociality in users. Assays could be developed to test AI on sample groups to measure its short- and long-term psychological impacts on users, data that is unfortunately largely missing at the present time. Perhaps, akin to FDA regulations on new drugs, new AI that is slated to be released to a wider public should be put through a battery of tests to show that, at the very least, it does no psychological harm. Drug companies are required to show extensive safety data before releasing a product. AI companies currently are not. It is in this space that government regulation of AI makes sense to us.

Others have made claims in the name of ethics about regulating characteristics of AI; however, these suggestions seem outdated. According to Bryson (2010), robots should be "slaves"—this does not mean that we should make robots slaves, but rather, we should keep them at a simpler developmental level by not giving them characteristics that might enable people to view them as anything other than owned and created by humans for humans. Bryson claims that it would be immoral to create a robot that can feel emotions like pain. Metzinger (2021) called for a ban on development of AI that could be considered sentient. AI advancement, however, continues in this direction. Calls for stopping the technological progress have not been effective. Relatively early in development of social actor AI, computer science researchers created benchmarks for human likeness to enable people to create more humanlike AI (Kahn et al., 2007). That human likeness has increased since. Our proposal has less to do with regulating how advanced or how humanlike AI becomes, and more to do with regulating how AI impacts the psychology of users by providing a model for prosocial behavior or by ignoring, confronting, or rectifying antisocial behavior.

Almost all discussion of regulating AI centers around its potential for harm. We will end this article by noting the enormous potential for benefit, especially in light of AI's guaranteed permanence in our present and future. Social AI is increasingly similar to humans in that it can engage in humanlike discourse, appear humanlike, and impact our social attitudes and interactions. Yet, social AI differs from humans in at least one significant way: it does not experience social or emotional fatigue. The opportunity to practice prosocial behavior is endless. For example, a chatbot will not grow tired and upset if you need to constructively work through a conflict with it. Neither will a chatbot disappear in the middle of a conversation when you are experiencing sadness or hurt and are in need of a friend. Social actor AI can both provide support and model prosocial behavior by remaining polite and present. Chatbots like WoeBot help users work through difficult issues by asking questions in the style of cognitive behavioral therapy (Fitzpatrick et al., 2017). Much like the benefits of journaling (Pennebaker, 1997, 2004), this human-chatbot engagement guides the user to make meaning of their experiences. It is worth noting that people who feel isolated or have experienced social rejection or social frustration may be a significant source of political and social disruption in today's world. If a universally available companion bot could boost their sense of social well-being and allow them to improve their social interaction skills through practice, that tool could make a sizable contribution to society. If AI is regulated such that it encourages people to treat it in a positive, pro-social way, and if carry-over effects are real, then AI becomes a potential source of enormous social and psychological good in the world.

If we are to effectively tackle the ever-growing issue of what to do in response to the surge of AI in our world, we cannot continue to point out only the ways in which it is harmful. AI is here to stay, and therefore we should be pragmatic with our approach. By understanding the ways in which interactions with AI can be both positive and negative, we can start to mitigate the bad by replacing it with the good.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RG: Conceptualization, Funding acquisition, Investigation, Resources, Validation, Writing – original draft, Writing – review & editing. MG: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. RG is funded by the National Science Foundation Graduate Research Fellowship Program. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. KB0013612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abubshait, A., and Wiese, E. (2017). You look human, but act like a machine: agent appearance and behavior modulate different aspects of human-robot interaction. *Front. Psychol.* 8:1393. doi: 10.3389/fpsyg.2017.01393

Akst, D. (2023). Should robots with artificial intelligence have moral or legal rights? WSJ. Available at: https://www.wsj.com/articles/robots-ai-legal-rights-3c47ef40

Alexander, F., and French, T. M. (1946). *Psychoanalytic Therapy: Principles and Application*. New York: Ronald Press.

Anderson, C. A., and Bushman, B. J. (2002). Human aggression. *Annu. Rev. Psychol.* 53, 27–51. doi: 10.1146/annurev.psych.53.100901.135231

Appel, M., Izydorczyk, D., Weber, S., Mara, M., and Lischetzke, T. (2020). The uncanny of mind in a machine: humanoid robots as tools, agents, and experiencers. *Comput. Hum. Behav.* 102, 274–286. doi: 10.1016/j.chb.2019.07.031

Appel, J., Von Der Pütten, A., Krämer, N. C., and Gratch, J. (2012). Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Adv. Hum. Comput. Interact.* 2012, 1–10. doi: 10.1155/2012/324694

Baars, B. J. (1997). In the Theater of Consciousness.

Bandura, A. (1965). Influence of models' reinforcement contingencies on the acquisition of imitative responses. *J. Pers. Soc. Psychol.* 1, 589–595. doi: 10.1037/h0022070

Bandura, A. (1977). *Social Learning Theory*. Englewood Cliffs, N.J.: Prentice Hall.

Banks, J. (2019). Theory of mind in social robots: replication of five established human tests. *Int. J. Soc. Robot.* 12, 403–414. doi: 10.1007/s12369-019-00588-x

Bastian, B., and Haslam, N. (2010). Excluded from humanity: the dehumanizing effects of social ostracism. *J. Exp. Soc. Psychol.* 46, 107–113. doi: 10.1016/j.jesp.2009.06.022

Beneteau, E., Boone, A., and Wu, Y., Kientz, J.A., Yip, J., and Hiniker, A. (2020). "Parenting with Alexa: exploring the introduction of smart speakers on family dynamics" in *Proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA. 1–13.

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., and Hiniker, A. (2019). "Breakdowns between families and Alexa" in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. Association for Computing Machinery, New York, NY, USA. 14.

Brandtzæg, P. B., Skjuve, M., and Følstad, A. (2022). My AI friend: how users of a social chatbot understand their human–AI friendship. *Hum. Commun. Res.* 48, 404–429. doi: 10.1093/hcr/hqac008

Broadbent, E., Kumar, V., Li, X., Sollers, J. J., Stafford, R., MacDonald, B. A., et al. (2013). Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PLoS One* 8:e72589. doi: 10.1371/journal.pone.0072589

Bryson, J. J. (2010). "Robots Should be Slaves", in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Ed. Yorick Wilks John Benjamins Publishing Company eBooks, 63–74.

Burleigh, T., Schoenherr, J. R., and Lacroix, G. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Comput. Hum. Behav.* 29, 759–771. doi: 10.1016/j.chb.2012.11.021

Bushman, B. J. (2002). Does venting anger feed or extinguish the flame? Catharsis, rumination, distraction, anger, and aggressive responding. *Personal. Soc. Psychol. Bull.* 28, 724–731. doi: 10.1177/0146167202289002

Bernstein, D., and Crowley, K. (2008). Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *Journal of the Learning Sciences* 17, 225–247. doi: 10.1080/10508400801986116

Chalmers, D. J. (1996). *Facing Up to the Problem of Consciousness*. The MIT Press eBooks.

Chalmers, D. J. (2023). Could a large language model be conscious? arXiv [Preprint]. doi: 10.48550/arxiv.2303.07103

Cheetham, M., Suter, P., and Jäncke, L. (2014). Perceptual discrimination difficulty and familiarity in the Uncanny Valley: more like a "Happy Valley". *Front. Psychol.* 5:1219. doi: 10.3389/fpsyg.2014.01219

Ciardo, F., De Tommaso, D., and Wykowska, A (2021). "Effects of erring behavior in a human-robot joint musical task on adopting intentional stance toward the iCub robot" in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. Vancouver, BC, Canada, 698–703.

Ciardo, F., De Tommaso, D., and Wykowska, A. (2022). Joint action with artificial agents: human-likeness in behaviour and morphology affects sensorimotor signaling and social inclusion. *Comput. Hum. Behav.* 132:107237. doi: 10.1016/j.chb.2022.107237

Croes, E., and Antheunis, M. L. (2020). Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *J. Soc. Pers. Relat.* 38, 279–300. doi: 10.1177/0265407520959463

Denzler, M., and Förster, J. (2012). A goal model of catharsis. *Eur. Rev. Soc. Psychol.* 23, 107–142. doi: 10.1080/10463283.2012.699358

Doerig, A., Schurger, A., and Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* 12, 41–62. doi: 10.1080/17588928.2020.1772214

Dollard, J., and Miller, N. E. (1950). *Personality and Psychotherapy*. New York: McGraw-Hill.

Dubosc, C., Gorisse, G., Christmann, O., Fleury, S., Poinsot, K., and Richir, S. (2021). Impact of avatar facial anthropomorphism on body ownership, attractiveness and social presence in collaborative tasks in immersive virtual environments. *Comput. Graph.* 101, 82–92. doi: 10.1016/j.cag.2021.08.011

Duffy, B. (2008). Fundamental issues in affective intelligent social machines. *Open Artif. Intellig. J.* 2, 21–34. doi: 10.2174/1874061800802010021

Edwards, A., Edwards, C., Westerman, D., and Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Comput. Hum. Behav.* 90, 308–314. doi: 10.1016/j.chb.2018.08.042

Etzrodt, K., and Engesser, S. (2021). Voice-based agents as personified things: assimilation and accommodation as equilibration of doubt. *Hum. Machine Commun. J.* 2, 57–79. doi: 10.30658/hmc.2.3

Etzrodt, K., Gentzel, P., Utz, S., and Engesser, S. (2022). Human-machine-communication: introduction to the special issue. *Publizistik* 67, 439–448. doi: 10.1007/s11616-022-00754-8

Eyssel, F. A., and Pfundmair, M. (2015). "Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: A case study with a zoomorphic robot" in *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication*.

Ferrari, F., Paladino, M. P., and Jetten, J. (2016). Blurring human–machine distinctions: anthropomorphic appearance in social robots as a threat to human distinctiveness. *Int. J. Soc. Robot.* 8, 287–302. doi: 10.1007/s12369-016-0338-y

Festerling, J., Siraj, I., and Malmberg, L. E. (2022). Exploring children's exposure to voice assistants and their ontological conceptualizations of life and technology. *AI & Soc.* doi: 10.1007/s00146-022-01555-3

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to Young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health* 4:e7785. doi: 10.2196/mental.7785

Fox, J., Ahn, S. J., Janssen, J., Yeykelis, L., Segovia, K. Y., and Bailenson, J. N. (2014). Avatars versus agents: a meta-analysis quantifying the effect of agency on social influence. *Hum. Comput. Interact.* 30, 401–432. doi: 10.1080/07370024.2014.921494

Frith, C. D. (2002). Attention to action and awareness of other minds. *Conscious. Cogn.* 11, 481–487. doi: 10.1016/s1053-8100(02)00022-3

Fu, Y., Michelson, R., Lin, Y., Nguyen, L. K., Tayebi, T. J., and Hiniker, A. (2022). Social emotional learning with conversational agents. *Proc. ACM Interact. Mobile Wearable Ubiquit. Technol.* 6, 1–23. doi: 10.1145/3534622

Garg, R., and Sengupta, S. (2020). He is just like me. *Proc. ACM Interact. Mobile Wearable Ubiquit. Technol.* 4, 1–24. doi: 10.1145/3381002

Gong, L., and Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: human identity and consistency preference. *Hum. Commun. Res.* 33, 163–193. doi: 10.1111/j.1468-2958.2007.00295.x

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

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

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. New York, NY: Oxford University Press.

Guingrich, R., and Graziano, M. S. A. (2023). Chatbots as social companions: how people perceive consciousness, human likeness, and social health benefits in machines (arXiv:2311.10599). arXiv [Preprint]. doi: 10.48550/arXiv.2311.10599

Han, M. C. (2021). The impact of anthropomorphism on consumers' purchase decision in chatbot commerce. *J. Internet Commer.* 20, 46–65. doi: 10.1080/15332861.2020.1863022

Harley, T. A. (2021). *The Science of Consciousness*. Cambridge, UK: Cambridge University Press.

Harris, P. L. (1992). From simulation to folk psychology: the case for development. *Mind Lang.* 7, 120–144. doi: 10.1111/j.1468-0017.1992.tb00201.x

Haslam, N., and Loughnan, S. (2014). Dehumanization and infrahumanization. *Annu. Rev. Psychol.* 65, 399–423. doi: 10.1146/annurev-psych-010213-115045

Hayashi, Y., and Miwa, K. (2009). "Cognitive and emotional characteristics of communication in human-human/human-agent interaction" in *Proceedings of the 13th International Conference on Human-Computer Interaction. Part III: Ubiquitous and Intelligent Interaction*. Springer Science & Business Media, 267–274.

Heyselaar, E., and Bosse, T. (2020). "Using Theory of Mind to Assess Users' Sense of Agency in Social Chatbots," in *Chatbot Research and Design*. Eds. A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, and P. B. Brandtzaeg. Vol. 11970 (Springer International Publishing), 158–169.

Hiniker, A., Wang, A., Tran, J., Zhang, M. R., Radesky, J., Sobel, K., et al. (2021). Can Conversational Agents Change the Way Children Talk to People? in: *IDC '21: Proceedings of the 20th Annual ACM Interaction Design and Children Conference*, 338–349.

Ho, A. S., Hancock, J., and Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *J. Commun.* 68, 712–733. doi: 10.1093/joc/jqy026

Hoyt, C. L., Blascovich, J., and Swinth, K. R. (2003). Social inhibition in immersive virtual environments. *Presence Teleoperat. Virtual Environ.* 12, 183–195. doi: 10.1162/10547603321640932

Jacobs, O., Gazzaz, K., and Kingstone, A. (2021). Mind the robot! Variation in attributions of mind to a wide set of real and fictional robots. *Int. J. Soc. Robot.* 14, 529–537. doi: 10.1007/s12369-021-00807-4

Kahn, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., et al. (2007). What is a human? *Interact. Stud.* 8, 363–390. doi: 10.1075/is.8.3.04kah

Kahn, P. H. Jr., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., et al. (2011). "The new ontological category hypothesis in human-robot interaction" in HRI'11. Association for Computing Machinery, New York, NY, USA. 159–160.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., et al. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Dev. Psychol.* 48, 303–314. doi: 10.1037/a0027033

Kätsyri, J., Förger, K., Mäkäräinen, M., and Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Front. Psychol.* 6:390. doi: 10.3389/fpsyg.2015.00390

Kawabe, T., Sasaki, K., Ihaya, K., and Yamada, Y. (2017). When categorization-based stranger avoidance explains the uncanny valley: a comment on MacDorman and Chattopadhyay (2016). *Cognition* 161, 129–131. doi: 10.1016/j.cognition.2016.09.001

Kim, D., Frank, M. G., and Kim, S. T. (2014). Emotional display behavior in different forms of computer mediated communication. *Comput. Hum. Behav.* 30, 222–229. doi: 10.1016/j.chb.2013.09.001

Kim, W., and Ryoo, Y. (2022). Hypocrisy induction: using chatbots to promote COVID-19 social distancing. *Cyberpsychol. Behav. Soc. Netw.* 25, 27–36. doi: 10.1089/cyber.2021.0057

Kim, Y., and Sundar, S. S. (2012). Anthropomorphism of computers: is it mindful or mindless? *Comput. Hum. Behav.* 28, 241–250. doi: 10.1016/j.chb.2011.09.006

Knobe, J., and Prinz, J. (2007). Intuitions about consciousness: experimental studies. *Phenomenol. Cogn. Sci.* 7, 67–83. doi: 10.1007/s11097-007-9066-y

Koch, C. (2019). The feeling of life itself: why consciousness is widespread but Can't be computed. Available at: https://openlibrary.org/books/OL29832851M/Feeling_of_Life_Itself

Konečni, V. (2016). The anger-aggression bidirectional-causation (AABC) model's relevance for dyadic violence, re-venge and catharsis. *Soc. Behav. Res. Pract. Open J.* 1, 1–9. doi: 10.17140/SBRPOJ-1-101

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS One* 3:e2597. doi: 10.1371/journal.pone.0002597

Krämer, N. C., Bente, G., Eschenburg, F., and Troitzsch, H. (2009). Embodied conversational agents: research prospects for social psychology and an exemplary study. *Soc. Psychol.* 40, 26–36. doi: 10.1027/1864-9335.40.1.26

Krämer, N., Bente, G., and Piesk, J. (2003a). The ghost in the machine. The influence of Embodied Conversational Agents on user expectations and user behavior in a TV/VCR application. ResearchGate. Available at: https://www.researchgate.net/publication/242273054_The_ghost_in_the_machine_The_influence_of_Embodied_Conversational_Agents_on_user_expectations_and_user_behaviour_in_a_TVVCR_application1

Krämer, N. C., Rosenthal-von der Pütten, A. M., and Hoffmann, L. (2015). "Social effects of virtual and robot companions" in *The Handbook of the Psychology of Communication Technology, Ch. 6* (John Wiley & Sons, Ltd.), 137–159.

Krämer, N. C., Tietz, B., and Bente, G. (2003b). "Effects of embodied Interface agents and their gestural activity" in *4th International Working Conference on Intelligent Virtual Agents*. Hamburg: Springer. 292–300.

Kupferberg, A., Glasauer, S., Huber, M., Rickert, M., Knoll, A., and Brandt, T. (2011). Biological movement increases acceptance of humanoid robots as human partners in motor interaction. *AI & Soc.* 26, 339–345. doi: 10.1007/s00146-010-0314-2

Küster, D., and Świderska, A. (2020). Seeing the mind of robots: harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes. *Int. J. Psychol.* 56, 454–465. doi: 10.1002/ijop.12715

Küster, D., Świderska, A., and Gunkel, D. J. (2020). I saw it on YouTube! How online videos shape perceptions of mind, morality, and fears about robots. *New Media Soc.* 23, 3312–3331. doi: 10.1177/1461444820954199

Lee, E. (2010). The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Comput. Hum. Behav.* 26, 665–672. doi: 10.1016/j.chb.2010.01.003

Lee, K. M., Jung, Y., Kim, J., and Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: the effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *Int. J. Hum. Comput. Stud.* 64, 962–973. doi: 10.1016/j.ijhcs.2006.05.002

Lee, S. A., and Liang, Y. (2016). The role of reciprocity in verbally persuasive robots. *Cyberpsychol. Behav. Soc. Netw.* 19, 524–527. doi: 10.1089/cyber.2016.0124

Lee, S. A., and Liang, Y. (2019). Robotic foot-in-the-door: using sequential-request persuasive strategies in human-robot interaction. *Comput. Hum. Behav.* 90, 351–356. doi: 10.1016/j.chb.2018.08.026

Lee, S., Ratan, R., and Park, T. (2019). The voice makes the Car: enhancing autonomous vehicle perceptions and adoption intention through voice agent gender and style. *Multimod. Technol. Interact.* 3:20. doi: 10.3390/mti3010020

Lew, Z., and Walther, J. B. (2022). Social scripts and expectancy violations: evaluating communication with human or AI Chatbot Interactants. *Media Psychol.* 26, 1–16. doi: 10.1080/15213269.2022.2084111

Loh, J., and Loh, W. (2023). Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds With Robots. transcript Verlag. Bielefeld, Germany.

Lucas, G. M., Gratch, J., King, A., and Morency, L. (2014). It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* 37, 94–100. doi: 10.1016/j.chb.2014.04.043

Luria, M., Sheriff, O., Boo, M., Forlizzi, J., and Zoran, A. (2020). Destruction, catharsis, and emotional release in human-robot interaction. *ACM Trans. Hum. Robot Interaction* 9, 1–19. doi: 10.1145/3385007

MacDorman, K. F., and Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146, 190–205. doi: 10.1016/j.cognition.2015.09.019

MacDorman, K. F., and Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interact. Stud.* 16, 141–172. doi: 10.1075/is.16.2.01mac

Marchesi, S., De Tommaso, D., Pérez-Osorio, J., and Wykowska, A. (2022). Belief in sharing the same phenomenological experience increases the likelihood of adopting the intentional stance toward a humanoid robot. *Technol Mind Behav* 3:11. doi: 10.1037/tmb0000072

Martini, M. C., Gonzalez, C., and Wiese, E. (2016). Seeing minds in others—can agents with robotic appearance have human-like preferences? *PLoS One* 11:e0146310. doi: 10.1371/journal.pone.0146310

McVee, M. B., Dunsmore, K., and Gavelek, J. R. (2005). Schema theory revisited. *Rev. Educ. Res.* 75, 531–566. doi: 10.3102/00346543075004531

Merritt, T. R. (2012). A failure of imagination: a failure of imagination: how and why people respond differently to human and computer team-mates. ResearchGate. Available at: https://www.researchgate.net/publication/292539389_A_failure_of_imagination_How_and_why_people_respond_differently_to_human_and_computer_team-mates

Metzinger, T. (2021). Artificial suffering: an argument for a global moratorium on synthetic phenomenology. *J. Artific. Intellig. Consciousness* 8, 43–66. doi: 10.1142/s270507852150003x

Mikropoulos, T. A., Misailidi, P., and Bonoti, F. (2003). Attributing human properties to computer artifacts: developmental changes in children's understanding of the animate-inanimate distinction. *Psychology* 10, 53–64. doi: 10.12681/psy_hps.23951

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy* 7, 33–35.

Nagel, T. (1974). What is it like to be a bat? *Philos. Rev.* 83:435. doi: 10.2307/2183914

Nass, C., and Brave, S. (2005). Wired for speech: How voice activates and advances the human-computer relationship. Boston Review: Boston, Massachusetts.

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

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers are social actors" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.

Nigam, M. K, and Klahr, D. (2000). "If robots make choices, are they alive?: Children's judgments of the animacy of intelligent artifacts" in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 22. Available at: https://escholarship.org/uc/item/6bw2h51d

O'Regan, J. K. (2012). How to build a robot that is conscious and feels. *Mind. Mach.* 22, 117–136. doi: 10.1007/s11023-012-9279-x

Opfer, J. E. (2002). Identifying living and sentient kinds from dynamic information: the case of goal-directed versus aimless autonomous movement in conceptual change. *Cognition* 86, 97–122. doi: 10.1016/s0010-0277(02)00171-3

Opfer, J. E., and Siegler, R. S. (2004). Revisiting preschoolers' living things concept: a microgenetic analysis of conceptual change in basic biology. *Cogn. Psychol.* 49, 301–332. doi: 10.1016/j.cogpsych.2004.01.002

Ortony, A., and Anderson, R. C. (1977). Definite descriptions and semantic memory. *Cogn. Sci.* 1, 74–83. doi: 10.1016/s0364-0213(77)80005-0

Pałka, P. (2023). AI, consumers & psychological harm (SSRN scholarly paper 4564997). Available at: https://papers.ssrn.com/abstract=4564997

Pankin, J. (2013). Schema theory and concept formation. Presentation at MIT, Fall. Available at: https://web.mit.edu/pankin/www/Schema_Theory_and_Concept_Formation.pdf

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychol. Sci.* 8, 162–166. doi: 10.1111/j.1467-9280.1997.tb00403.x

Pennebaker, J. W. (2004). *Writing to Heal: A Guided Journal for Recovering from Trauma and Emotional Upheaval*. Oakland, CA: New Harbringer Publications.

Pentina, I., Hancock, T., and Xie, T. (2023). Exploring relationship development with social chatbots: a mixed-method study of replika. *Comput. Hum. Behav.* 140:107600. doi: 10.1016/j.chb.2022.107600

Poushneh, A. (2021). Humanizing voice assistant: the impact of voice assistant personality on consumers' attitudes and behaviors. *J. Retail. Consum. Serv.* 58:102283. doi: 10.1016/j.jretconser.2020.102283

Powers, A., and Kiesler, S. (2006). "The advisor robot: tracing people's mental model from a robot's physical attributes" in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, Salt Lake City, USA. 218–225.

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/s0140525x00076512

Prinz, W. (2017). Modeling self on others: an import theory of subjectivity and selfhood. *Conscious. Cogn.* 49, 347–362. doi: 10.1016/j.concog.2017.01.020

Pütten, A. M. R. D., and Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. *Comput. Hum. Behav.* 36, 422–439. doi: 10.1016/j.chb.2014.03.066

Quadflieg, S., Ul-Haq, I., and Mavridis, N. (2016). Now you feel it, now you don't. *Interact. Stud.* 17, 211–247. doi: 10.1075/is.17.2.03qua

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018). Machine theory of mind. arXiv [Preprint]. doi: 10.48550/ARXIV.1802.07740

Rhim, J., Kwak, M., Gong, Y., and Gweon, G. (2022). Application of humanization to survey chatbots: change in chatbot perception, interaction experience, and survey data quality. *Comput. Hum. Behav.* 126:107034. doi: 10.1016/j.chb.2021.107034

Rickenberg, R., and Reeves, B. (2000). The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. *Lette. CHI* 2000, 49–56. doi: 10.1145/332040.332406

Roselli, C., Navare, U. P., Ciardo, F., and Wykowska, A. (2023). Type of education affects individuals' adoption of intentional stance towards robots: an EEG study. *Int. J. Soc. Robot.* 16, 185–196. doi: 10.1007/s12369-023-01073-2

Röska-Hardy, L. (2008). "Theory (Simulation Theory, Theory of Mind)", *in Encyclopedia of Neuroscience*. Eds. M. Binder, N. Hirokawa, U. Windhorst and H. Hirsch, Berlin/Heidelberg Germany: Springer eBooks, 4064–4067.

Ryland, H. (2021a). It's friendship, Jim, but not as we know it: a degrees-of-friendship view of human–robot friendships. *Mind. Mach.* 31, 377–393. doi: 10.1007/s11023-021-09560-z

Ryland, H. (2021b). Could you hate a robot? And does it matter if you could? *AI & Soc.* 36, 637–649. doi: 10.1007/s00146-021-01173-5

Scaife, M., and Van Duuren, M. V. (1995). Do computers have brains? What children believe about intelligent artifacts. *Br. J. Dev. Psychol.* 13, 367–377. doi: 10.1111/j.2044-835x.1995.tb00686.x

Seeger, A., and Heinzl, A. (2018). "Human versus machine: contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents" in *Lecture Notes in Information Systems and Organisation*, Eds. F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, and A. B. Randolph. Springer International Publishing. 129–139.

Severson, R. L., and Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Netw.* 23, 1099–1103. doi: 10.1016/j.neunet.2010.08.014

Severson, R. L., and Woodard, S. R. (2018). Imagining others' minds: the positive relation between children's role play and anthropomorphism. *Front. Psychol.* 9:2140. doi: 10.3389/fpsyg.2018.02140

Shank, D. B., North, M., Arnold, C., and Gamez, P. (2021). Can mind perception explain virtuous character judgments of artificial intelligence? *Technol Mind Behav* 2. doi: 10.1037/tmb0000047

Spence, P. R., Westerman, D., Edwards, C., and Edwards, A. (2014). Welcoming our robot overlords: initial expectations about interaction with a robot. *Commun. Res. Rep.* 31, 272–280. doi: 10.1080/08824096.2014.924337

Sproull, L., Subramani, M. R., Kiesler, S., Walker, J., and Waters, K. (1996). When the interface is a face. *Hum. Comput. Interact.* 11, 97–124. doi: 10.1207/s15327051hci1102_1

Srinivasan, V., and Takayama, L. (2016). "Help me please: robot politeness strategies for soliciting help from humans" in *CHI'16*. Association for Computing Machinery, New York, NY, USA. 4945–4955.

Stein, J., Appel, M., Jost, A., and Ohler, P. (2020). Matter over mind? How the acceptance of digital entities depends on their appearance, mental prowess, and the interaction between both. *Int. J. Hum. Comput. Stud.* 142:102463. doi: 10.1016/j.ijhcs.2020.102463

Stein, J., and Ohler, P. (2017). Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition* 160, 43–50. doi: 10.1016/j.cognition.2016.12.010

Sundar, S. S., and Nass, C. (2000). Source orientation in human-computer interaction. *Commun. Res.* 27, 683–703. doi: 10.1177/009365000027006001

Świderska, A., and Küster, D. (2018). Avatars in pain: visible harm enhances mind perception in humans and robots. *Perception* 47, 1139–1152. doi: 10.1177/0301006618809919

Ta, V. P., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., et al. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *J. Med. Internet Res.* 22:e16235. doi: 10.2196/16235

Tanibe, T., Hashimoto, T., and Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLoS One* 12:e0180952. doi: 10.1371/journal.pone.0180952

Taylor, J., Weiss, S. M., and Marshall, P. (2020). Alexa, how are you feeling today? *Interact. Stud.* 21, 329–352. doi: 10.1075/is.19015.tay

Teubner, T., Adam, M. T. P., and Riordan, R. (2015). The impact of computerized agents on immediate emotions, overall arousal and bidding behavior in electronic auctions. *J. Assoc. Inf. Syst.* 16, 838–879. doi: 10.17705/1jais.00412

Tharp, M., Holtzman, N. S., and Eadeh, F. R. (2016). Mind perception and individual differences: a replication and extension. *Basic Appl. Soc. Psychol.* 39, 68–73. doi: 10.1080/01973533.2016.1256287

Tononi, G. (2007). "The information integration theory of consciousness," *The Blackwell companion to consciousness*. Eds. M. Velmans and S. Schneider (Oxford: Blackwell), 287–299.

UNESCO (2020). Artificial intelligence and gender equality: Key findings of UNESCO's Global Dialogue—UNESCO Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000374174 (Accessed October 13, 2023).

UNESCO & EQUALS Skills Coalition Mark, W., Rebecca, K., and Chew, H. E. (2019). I'd blush if I could: Closing gender divides in digital skills through education—UNESCO Digital Library.

Velez, J. A., Loof, T., Smith, C. A., Jordan, J. M., Villarreal, J. A., and Ewoldsen, D. R. (2019). Switching schemas: do effects of mindless interactions with agents carry over to humans and vice versa? *J. Comput.-Mediat. Commun.*, 24, 335–352. doi: 10.1093/jcmc/zmz016

Vogeley, K., and Bente, G. (2010). "Artificial humans": psychology and neuroscience perspectives on embodiment and nonverbal communication. *Neural Netw.* 23, 1077–1090. doi: 10.1016/j.neunet.2010.06.003

Von Der Pütten, A. M., Reipen, C., Wiedmann, A., Kopp, S., and Krämer, N. C. (2009). "The impact of different embodied agent-feedback on users' behavior" in *Lecture Notes in Computer Science,* Eds. Z. Ruttkay, M. Kipp, A. Nijholt, and H. H. Vilhjálmsson, 549–551.

Wang, Q., Saha, K., Gregori, E., Joyner, D., and Goel, A. (2021). "Towards mutual theory of mind in human-ai interaction: how language reflects what students perceive about a virtual teaching assistant" in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 384, 1–14.

Ward, A. F., Olsen, A. S., and Wegner, D. M. (2013). The harm-made mind: observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychol. Sci.* 24, 1437–1445. doi: 10.1177/0956797612472343

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

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128.

Wilkenfeld, J. N., Yan, B., Huang, J., Luo, G., and Algas, K. (2022). "AI love you": linguistic convergence in human-chatbot relationship development. *Academy of Management Proceedings*, 17063. doi: 10.5465/AMBPP.2022.17063abstract

Worchel, P. (1957). Catharsis and the relief of hostility. *J. Abnorm. Soc. Psychol.* 55, 238–243. doi: 10.1037/h0042557

Xie, T., and Pentina, I. (2022). "Attachment theory as a framework to understand relationships with social Chatbots: a case study of Replika" in *Proceedings of the 55th Annual Hawaii International Conference on System Sciences*.

Yampolskiy, R. V. (2018). Artificial consciousness: an illusionary solution to the hard problem. *Reti Saperi Linguag.* 2, 287–318. doi: 10.12832/92302

Young, A. D., and Monroe, A. E. (2019). Autonomous morals: inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *J. Exp. Soc. Psychol.* 85:103870. doi: 10.1016/j.jesp.2019.103870

Zhan, J., Yu, S., Cai, R., Xu, H., Yang, Y., Ren, J., et al. (2021). The effects of written catharsis on anger relief. *PsyCh J.* 10, 868–877. doi: 10.1002/pchj.490

Zhou, Y., Fei, Z., He, Y., and Yang, Z. (2022). How Human–Chatbot Interaction Impairs Charitable Giving: The Role of Moral Judgment. *Journal of Business Ethics*, 178, 849–865. doi: 10.1007/s10551-022-05045-w

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 24 May 2024
DOI 10.3389/fpsyg.2024.1351425

OPEN ACCESS

EDITED BY
Paola Magnano,
Kore University of Enna, Italy

REVIEWED BY
Subhash Sagar,
Macquarie University, Australia
Ruining Jin,
China University of Political Science and Law, China

*CORRESPONDENCE
Chunhui Qi
✉ qchizz@126.com

RECEIVED 06 December 2023
ACCEPTED 13 May 2024
PUBLISHED 24 May 2024

CITATION
Zhang Z, Deng W, Wang Y and Qi C (2024) Visual analysis of trustworthiness studies: based on the Web of Science database. *Front. Psychol.* 15:1351425. doi: 10.3389/fpsyg.2024.1351425

Visual analysis of trustworthiness studies: based on the Web of Science database

Zhen Zhang[1,2], Wenqing Deng[1], Yuxin Wang[1] and Chunhui Qi[1]*

[1]Faculty of Education, Henan Normal University, Xinxiang, China, [2]Faculty of Education, Henan University, Kaifeng, China

Trustworthiness is the most significant predictor of trust and has a significant impact on people's levels of trust. Most trustworthiness–related research is empirical, and while it has a long history, it is challenging for academics to get insights that are applicable to their fields of study and to successfully transfer fragmented results into practice. In order to grasp their dynamic development processes through the mapping of network knowledge graphs, this paper is based on the Web of Science database and uses CiteSpace (6.2.R4) software to compile and visualize the 1,463 publications on trustworthy studies over the past 10 years. This paper aims to provide valuable references to theoretical research and the practice of Trustworthiness. The findings demonstrate that: over the past 10 years, trustworthiness-related research has generally increased in volume; trustworthiness research is concentrated in industrialized Europe and America, with American research findings having a bigger global impact; The University of California System, Harvard University, and Yale University are among the high-production institutions; the leading figures are represented by Alexander Todorov, Marco Brambilla, Bastian Jaeger, and others; the core authors are distinguished university scholars; however, the level of cooperation of the core author needs to be improved. The primary journal for publishing research on trustworthiness is the *Journal of Personality and Social Psychology* and *Biology Letters*. In addition, the study focuses on three distinct domains, involving social perception, facial clues, and artificial intelligence.

KEYWORDS

trustworthiness, visual analysis, mapping knowledge domains, Web of Science, CiteSpace

1 Introduction

Trust is an indispensable component of social life (Kennedy and Schweitzer, 2018; Bai et al., 2019; Zhu et al., 2022; Du et al., 2023) and serves as a lubricant for social integration (Yan and Wu, 2016; Milesi et al., 2023). Interpersonal trust is the cornerstone of social interaction and is crucial for society to function properly (Ścigała et al., 2020). However, the factors that arouse trust in others have predominantly not been investigated (Bellucci et al., 2019; Bennett, 2023). Trust is formed up of two elements, trust intention and trust belief, where trust belief is the perceived trustworthiness of others (Kim et al., 2004), including the ability, benevolence, and integrity of others (Mayer et al., 1995). Trustworthiness is the tendency of a trustee to meet the implicit or explicit positive expectations of others for a particular behavior, which reflects the degree to which a trustee is trustworthy (Levine et al., 2018). Trustworthiness is evaluated as (or lacking) the motive for lying as a proximal

antecedent variable of trust (Mayer et al., 1995). It is the most significant predictor of trust (Tomlinson et al., 2020) and is perceived as a social norm (Bicchieri et al., 2011). It is the basis for well–functioning interpersonal relationships and usually affects people's levels of trust. For instance, Van't Wout and Sanfey discovered that people with high levels of trustworthiness are more likely to gain their peers' trust and engage in collaboration than those with low levels (van 't Wout and Sanfey, 2008).

We are frequently forced to make trust decisions in relationships but also occasionally face situations involving trust violations. The choice of whether to trust others in interpersonal interactions can be considered a trust decision (Radke et al., 2018). Correctly trusting others can yield enormous rewards, while wrongly trusting others can have serious consequences. The collapse of trust connections frequently results in severe economic, emotional, and social costs for people (Bottom et al., 2002); however, trustworthiness may influence this outcome. The expression "breach of trust" means when a party's trust intention or belief decreases because of a trust policy being violated (Kim et al., 2009). A breach of trust may result in a variety of adverse effects, such as negative emotional, cognitive, and behavioral implications (Shu et al., 2021; van der Werff et al., 2023); acts of retaliation (Yan and Wu, 2016); a decreased sense of justice (Kennedy and Schweitzer, 2018); and the breakdown of bilateral cooperation (Bottom et al., 2002). This paper reviews the research in the field, considering the major effect that trustworthiness has on individuals' ability to assess the truth of assertions.

Trustworthiness has been studied for decades, but they are mostly empirical studies (e.g., Poon, 2013; Lleó de Nalda et al., 2016; Reimann et al., 2022). The review literature is few, a relatively limited amount of research has been quantitatively analyzed. So, it is impossible to fully describe the status and trend of trustworthiness research. This makes it difficult for researchers to gain insights that apply to their specific research area and to effectively translate fragmented research findings into practice. Therefore, this study intends to utilize bibliometric methods to collate and summarize previous studies on trustworthiness. Meanwhile, CiteSpace software is used to analyze the trustworthiness literature which collects in the Web of Science's core database over the past 10 years. With the aim of helping the researchers grasp changing trends and structures within relevant areas of research, references for an in-depth examination of the scenario that is currently in place and for cutting-edge dynamics and prospective trends in the field.

Bibliometrics is a branch of information science that has grown into one of the most active fields in the field of worldwide book intelligence. It reflects the trend of contemporary discipline quantification (Qiu et al., 2003). Quantitative analysis from the perspective of bibliometrics can summarize the development status of a research field more objectively. Nevertheless, by means of data mining, processing and measurement or mapping, CiteSpace can be used to graphically express knowledge frameworks, structures, interactions, intersections, derivations, and other internal connections (Liu et al., 2009). To guarantee the highest level of data display accuracy, the CiteSpace software has a sophisticated data processing system and strong visualization tools that include both structural and temporal indications. Researchers frequently use these indicators to carry out in-depth assessments and analyses of research frontiers and hotspots within domains, allowing them to promptly comprehend the most recent advancements and development patterns. One of the core functions of this software is detection and analysis of the research frontier and knowledge relationship (Jia et al., 2019), as a result, the CiteSpace knowledge graph has gained popularity in a variety of scientific fields for drafting literature reviews because of its benefits. By analyzing the status, hotspots and trends of trustworthiness research, researchers can benefit from the integrated and fragmented knowledge. The root causes and the latest development status can be understood. In addition, this way can also enrich trustworthy study contents and more easily apply researchers' findings to areas of interest to support them in assessing new directions for future research.

It is worth noting that the majority of the papers are qualitative-based, and previous researchers have conducted quantitative studies in different ways around different topics of trustworthiness, such as using meta-analysis (e.g., Yang and Beatty, 2016; Travers et al., 2019; Siddique et al., 2022) to quantify the extent, breadth, and role of age-related confidence differences (Bailey and Leon, 2019) and to evaluate the role of almond nuclei in facial trustworthiness treatment (Santos et al., 2016). Propose a paradigm change to advise on developing trustworthiness through ethical public health practices (Best et al., 2021). Rely on 79 peer-reviewed quantitative empirical studies spanning more than two decades to demonstrate the complexity of trust in a global homeschool context (Shayo et al., 2021). Investigate trustworthiness among human machines (Song and Luximon, 2020), and research trustworthiness using rooted theoretical techniques (e.g., Cheer et al., 2015; Filieri, 2016).

However, several study themes in the field of trustworthiness research complicate existing research, making it difficult to adequately reflect the status quo, research hotspots, and evolutionary tendencies. The current literature lacks a comprehensive study strategy, resulting in significant variation in the operability of existing studies (Siuda et al., 2022), preventing meaningful comparisons of findings as well as sufficient quantitative analysis. To acquire a thorough grasp of the current state and growth of trustworthiness studies, as well as to diminish their subjectivity, we undertake a comprehensive assessment using knowledge graph analysis, with the goal of depicting the area comprehensively and methodically. The approach of literature metrology allows scholars to reflect the state and substance of trustworthiness studies, highlight the development trajectory of trustworthiness research, increase their grasp of the field's evolution, and identify new directions more directly. On this basis, the study uses 1,463 pieces of relevant literature, quantitative literature measurement analysis, and trustworthiness-related studies to tackle the following research questions:

1. Which authors and journals are regularly referred to in trustworthiness studies, acting as a jumping–off point to find high–impact research in the field?
2. What is the volume distribution of trustworthiness by time and region?
3. What is the main field of trustworthiness research?
4. Which research areas indicate expectations for the future?

2 Methods

We set the thematic term to trustworthy or trustworthiness to collect effective and comprehensive objective literature. We applied certain limits before searching for subjects. First, we chose the Science

Citation Index Expanded (SCI-Expanded) and Social Sciences Citation index (SSCI) from the Core database of the Web of Science as our research platform. It is the largest, complete database of academic information covering disciplines, with a wide range of time, quantity, and quality. Second, there have been fewer trustworthiness-related studies between the creation of the Web of Science database in 1985 and 2013, with a total of 24 pertinent pieces of literature that are not analytically reliable. The data from the past decade is more current and representative, providing a better reflection of the current academic field's development trends and hotspots. Consequently, the relevant documentation published for 10 the past years was selected as the object of analysis, with the time range set to be from 30 July 2013 to 30 July 2023. Finally, we examine trustworthy as a psychological or perceptual value, referring to an individual's or an organization's characteristics or traits that inspire trust in them (Mayer et al., 1995). So, trustworthy in our study is a psychological trait, the literature type was set as article or review paper, and the literature category was limited to psychology, including multidisciplinary, social, applied, clinical, developmental, experimental, educational, biological, and mathematical psychology. A total of 1,463 studies were retrieved (see Figure 1).

The 1,463 works acquired by the study were visualized using an assortment of literature-determining and content-evaluation methods. First, the relevant documentation was obtained in pure text form from the Web of Science core collection database. Second, the visualization analysis software CiteSpace (6.2. R4) (hereinafter referred to as CiteSpace) was utilized to analyze the node type, which includes countries, authors, institutions, journals, and keywords. The time slice was set to 2 years. Third, we analyzed the results of the data analysis and relevant documentation, and the selected content was pruned by pathfinder to yield the corresponding knowledge maps.

3 Results

3.1 Spatiotemporal distribution

3.1.1 Annual publication volume analysis

The number of annual publications can reflect the development trend of a certain research field. Within the scope of retrieval, the annual amount of trustworthiness is shown in Figure 2, showing a rising trend. The development process in this field can be categorized into three stages: the phase of gradual advancement (pre-2014), the phase of rapid progression (2014–2015), and the phase of sustained growth (post-2015). We can discover that 2015 is a turning point in terms of items published by year. Before 2015, trustworthiness research reached a low level and continued in the enlightenment phase, this suggests that the study of trustworthiness is just beginning. After 2015, the volume of submissions significantly increased. Trustworthiness studies decline slightly in 2020 but then quickly recover and reach a new level in 2022 (236 articles). The publication of literature has varied over the past 10 years; instead, overall, the field of trustworthiness-related studies is moving forward. According to the

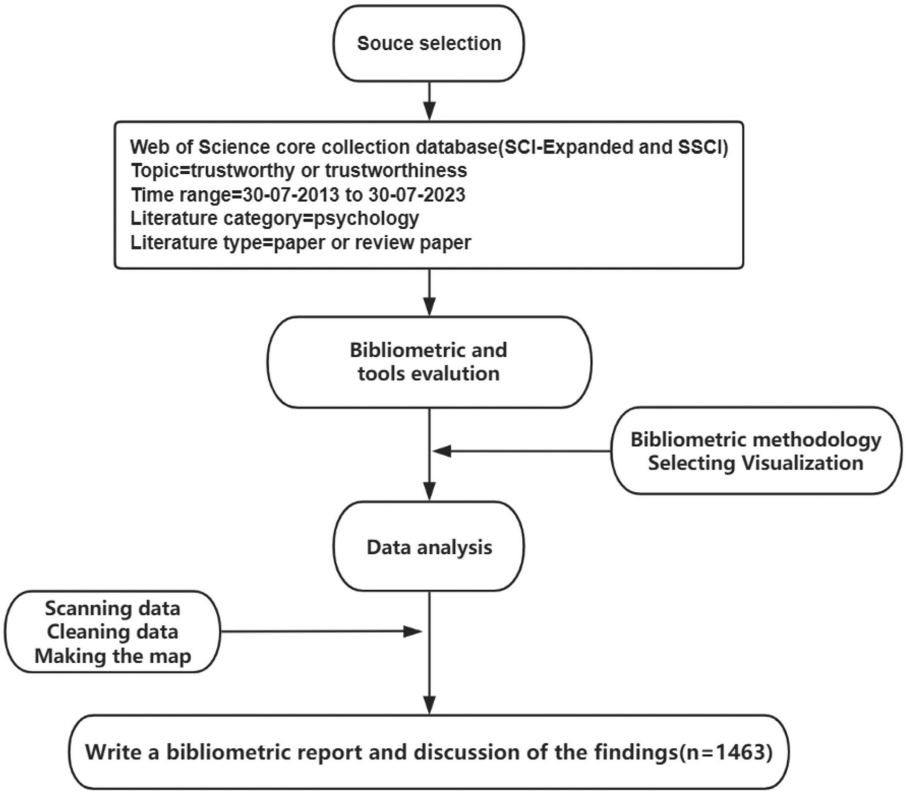


FIGURE 1
Process of making the map.

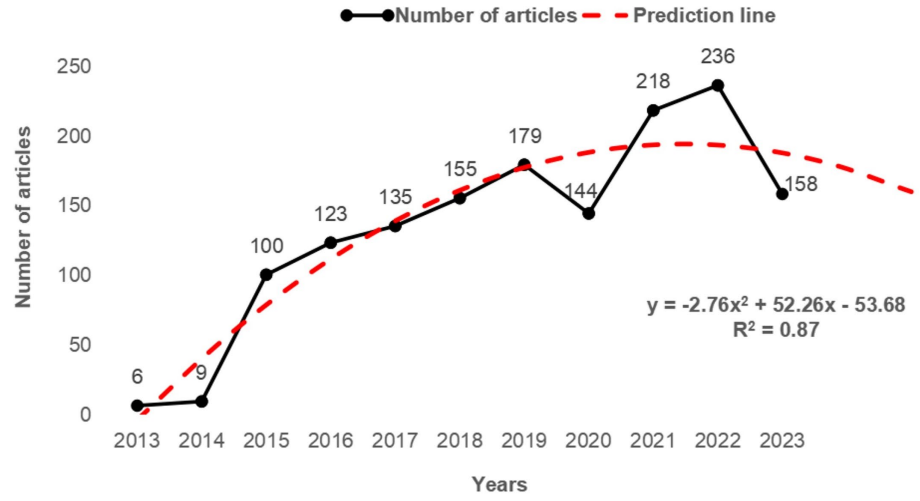


FIGURE 2
Annual publication volume.

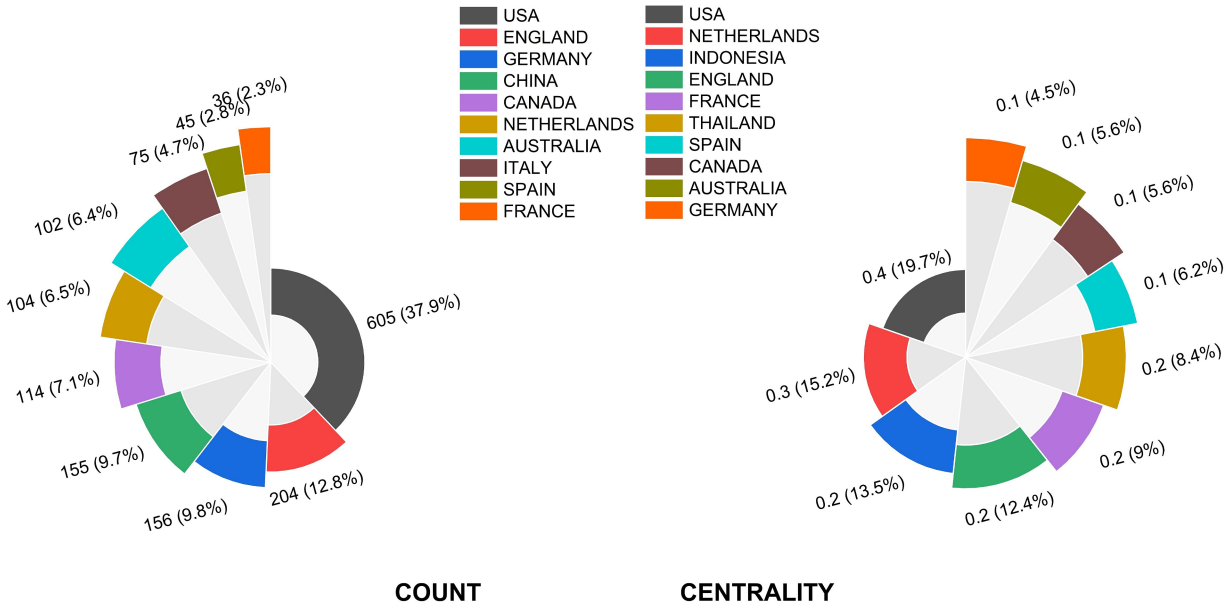


FIGURE 3
Chart of publication frequency of countries.

polynomial fitting curve, trustworthiness studies are expected to remain at a more stable level for the next 2 years.

3.1.2 Country analysis

The top 10 countries in terms of articles and the value of trustworthiness research are presented in Figure 3. The United States has the most documents and the highest degree of centrality within the search area, indicating that it has the closest academic research relationship with other countries and contributed considerably to research innovation, which had a major effect. Despite the importance of trust, scholars have not given it much consideration for a long time. This hush did not break until the 1950s, when American psychologist M. Deutsch conducted the first experimental investigation of trust in the Prisoner's Dilemma (Deutsch, 1958). Subsequently, many scholars in psychology, economics, sociology, political science, and other disciplines began to conduct in-depth research on trust issues from their own perspectives, resulting in an increase in trust research abroad and the formation of some relatively systematic trust research theories. This may be the reason why the United States is the highest degree of centrality country. The United Kingdom, Germany, China, Canada, the Netherlands, Australia, Italy, Spain, France, and other countries have high levels of production. The top three countries with the highest centrality are the United States, the Netherlands, and Indonesia. Germany and China publish the same number of papers, but their respective centralities are only 0.08 and 0.05, respectively, indicating that both have weak trustworthiness and should boost it.

3.2 Subject of publication

3.2.1 Issuing institutions analysis

Table 1 lists the top 10 institutions based on count, centrality, and burst value. The top three universities in terms of volume are the University of California System, the University of London, and the N8 Research Partnership, and the top 10 research institutions in terms of publication volume published 354 articles. Harvard University, the University of California System, and the State University System of Florida are the top three universities, accounting for 24% of the total literature, with centrality values of 0.19, 0.15, and 0.15, respectively, all three universities are from the United States. The authors and journals together further confirm the important role and status of the United States in the field of trustworthiness research. The top three universities in terms of burst values are Yale University, the University of Toronto, and Northwestern University, with 4.17, 3.89, and 3.09, respectively.

3.2.2 Cited journals analysis

CiteSpace provides an illustration of the year and name of the cited journal with the size and color of the "Year Wheel." The cited journal network knowledge graph involves a total of 241 nodes and 399 connections, with a density of 0.0138 within the criteria of the search (Figure 4). Evaluating the significant study results centered around trustworthiness becomes simpler through the analysis of these academic journals. As a result, Table 2 includes statistics for the 10 most common, centralized, and explosive publications. J Pers Soc Psychol was the most frequently cited journal (863 times), followed by Psychol Sci (724 times) and Psychol Bull (592 times). The journal with the highest centralization was J Pers Soc Psychol (0.43), followed by Psychol Sci (0.33) and J Appl Psychol (0.19). The most significant growth was in Biol Letters (12.27), followed by Thesis (9.50) and Nat Hum Behav (9.06).

Journal of Personality and Social Psychology mainly includes the empirical research reports related to personality and social psychology. Cottrell published an article entitled What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics in this journal had been cited more than 211 times. The article states that trustworthiness is considered extremely important for all the interdependent others in different measures of trait importance and different groups and relationships (Cottrell et al., 2007). Biology Letters is a professional bio-journal published by Royal Soc Publisher. Rhodes et al. (2012) published an article titled Women can judge sexual unfaithfulness from unfamiliar men's faces, which examines whether sexual trust (loyalty) can be accurately judged from the face of a stranger of the opposite sex. It concludes that for women, there is some evidence of judging sexual loyalty from their faces and further demonstrates that face perception appropriately adjusts to the signals of mate quality.

3.2.3 Author analysis

A network density of 0.0081, 249 nodes, and 249 connections are present. Only the authors with the highest links are shown in Figure 5. According to the figure, a crucial collaborative group has developed in the field of trustworthiness research with high-production authors like Sutherland, who primarily studies the detection of facial clues to trustworthiness. His most popular article is that social inferences from faces: ambient images generate a three-dimensional model, which developed and validated a 3D model (approachability, dominance and youthful-attractiveness) through 3 experiments, and studying two-dimensional valence or trustworthiness through a dominance model of face social inference. What's more, his findings highlight both the utility of the original trustworthiness and dominance dimensions and the need to utilize various facial stimuli, as well as further highlight the importance of youth and attractiveness perception in facial assessment (Sutherland et al., 2013).

As leaders in the field of research, high-production authors or core authors not only control the field's current research hotspots and directions but also influence the direction of subsequent research (see Tables 3, 4). Based on the number of submissions, the top three authors are Todorov A, Sutherland CAM, and Evans AM. The first three authors in burst value are Brambilla M, Tipper SP, and Rhodes G, with values of 3, 2.41, and 2.07, respectively. Brambilla M, Jaeger B, Masi M, and Mattavelli S are the most dynamic authors in the last 3 years when the number of citations is considered the number of citations of an article in other author references after publication. Todorov A, Oosterhof NN, and Willis J are the top three most cited authors. Jaeger B, Ma DS, and Glaeser EL are the top three most triggered authors, with values of 12.39, 10.22, and 8.96, respectively. The articles produced by the individuals mentioned above played a vital, pivotal, or revolutionary role in the research on trustworthiness.

One of the authors who is frequently cited, Todorov A, also emphasized the trustworthiness of faces, linking them to emotions

TABLE 1 Institutions distribution by count, centrality and burst.

Institutions	Count	Institutions	Centrality	Institutions	Burst
University of California System	52	Harvard University	0.19	Yale University	4.17
University of London	49	University of California System	0.15	University of Toronto	3.89
N8 Research Partnership	43	State University System of Florida	0.15	Northwestern University	3.09
Harvard University	37	Columbia University	0.15	University of Oxford	3.04
White Rose University Consortium	34	University of York—UK	0.11	University of Cologne	2.70
University System of Ohio	31	University of London	0.10	Maastricht University	2.39
University of York—UK	29	University of California Los Angeles	0.10	Duke University	2.37
Tilburg University	27	University of North Carolina	0.09	Radboud University Nijmegen	2.36
University of Milano-Bicocca	26	University of Western Australia	0.09	University of Cambridge	2.33
State University System of Florida	26	CIVIS	0.09	Renmin University of China	2.26

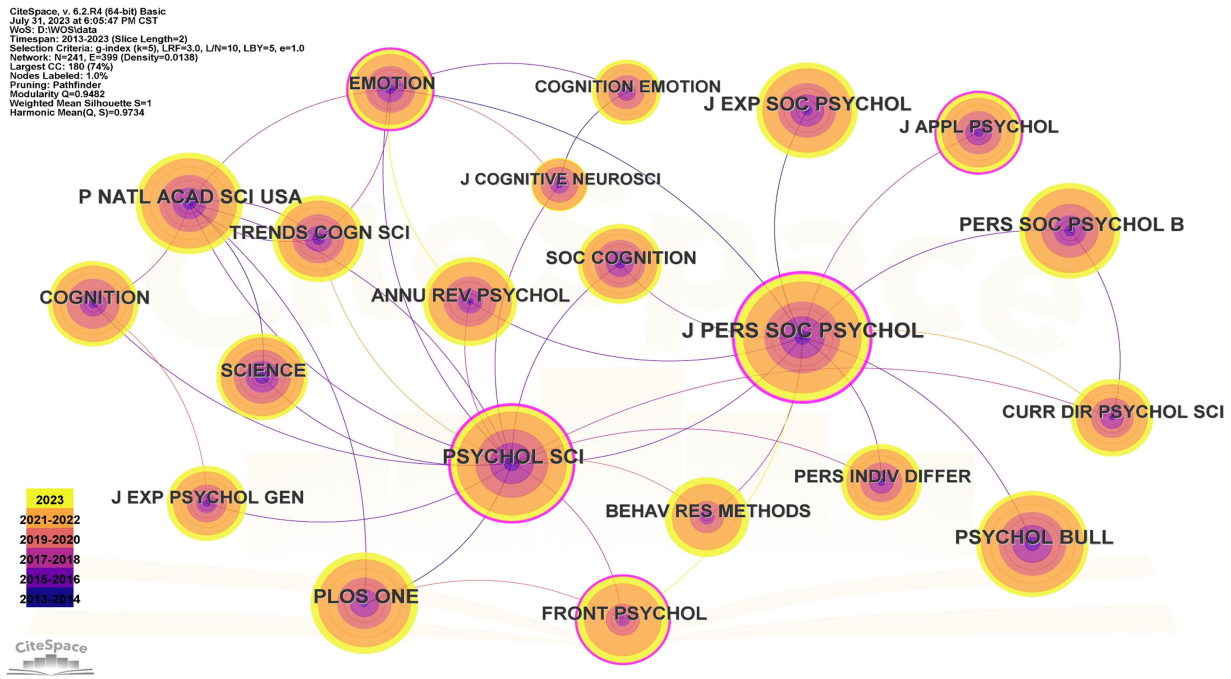


FIGURE 4
Cooperative network diagram of cited journals.

TABLE 2 Distribution of cited journals.

Journals	Count	Journals	Centrality	Journals	Burst
J Pers Soc Psychol	863	J Pers Soc Psychol	0.43	Biol Letters	12.27
Psychol Sci	724	Psychol Sci	0.33	Thesis	9.50
Psychol Bull	592	J Appl Psychol	0.19	Nat Hum Behav	9.06
PLoS One	556	Organ Behav Hum Dec	0.16	Psychol Med	9.02
P Natl Acad Sci USA	555	Comput Hum Behav	0.16	J Vision	8.98
J Exp Soc Psychol	492	Front Psychol	0.15	Nat Commun	8.81
Pers Soc Psychol B	484	Acad Manage J	0.13	Soc Sci Med	8.65
Annu Rev. Psychol	431	Emotion	0.12	Cortex	8.28
Trends Cogn Sci	416	Cognition	0.08	J Educ Meas	8.04
Cognition	402	Neuropsychologia	0.08	J Conflict Resolut	7.79

and neurons and studying how people perceive, evaluate, and absorb the social world. The author with the highest burst value, Brambilla M, also studied face trustworthiness. However, Brambilla showed that the facial aspect ratio and the tone of the voice can influence social perception. The research focused on the effect of auditory and visual cues on facial trustworthiness. The author Jaeger B, with the highest cited bursts, studied the effects of facial trustworthiness cues on social decision-making and social interactions. This indicates that research on trustworthiness focuses mainly on facial trustworthiness.

3.3 Thematic characteristics

3.3.1 Keyword analysis

Keywords act as both an overview of the topic and the thesis's main point of view. The amount of attention within an area of study can be seen by high keyword frequency and centralization. It may reflect advanced research methods, hot issues that need to be addressed, or academic topics that interest researchers over a given period. Within the scope of retrieval, there are 270 nodes and 340 links in the knowledge map of the keyword network, with a density of 0.0094 (see Figure 6). The top 10 keywords for frequency and centralization are listed in Table 5. The three keywords with the highest frequency are perception, trust, and trustworthiness, with frequencies of 387, 206, and 161, respectively. Competence, model, and facial expressions were the top three in terms of centrality, at 0.11, 0.10, and 0.09, respectively. Combining high-frequency keywords with associated literatures reveals that trustworthiness research focuses primarily on anex-dependent and post-dependent variables such as ability, facial expression, first impression, and So On. These keywords are closely related to other keywords around, which is more important and has greater influence in the research.

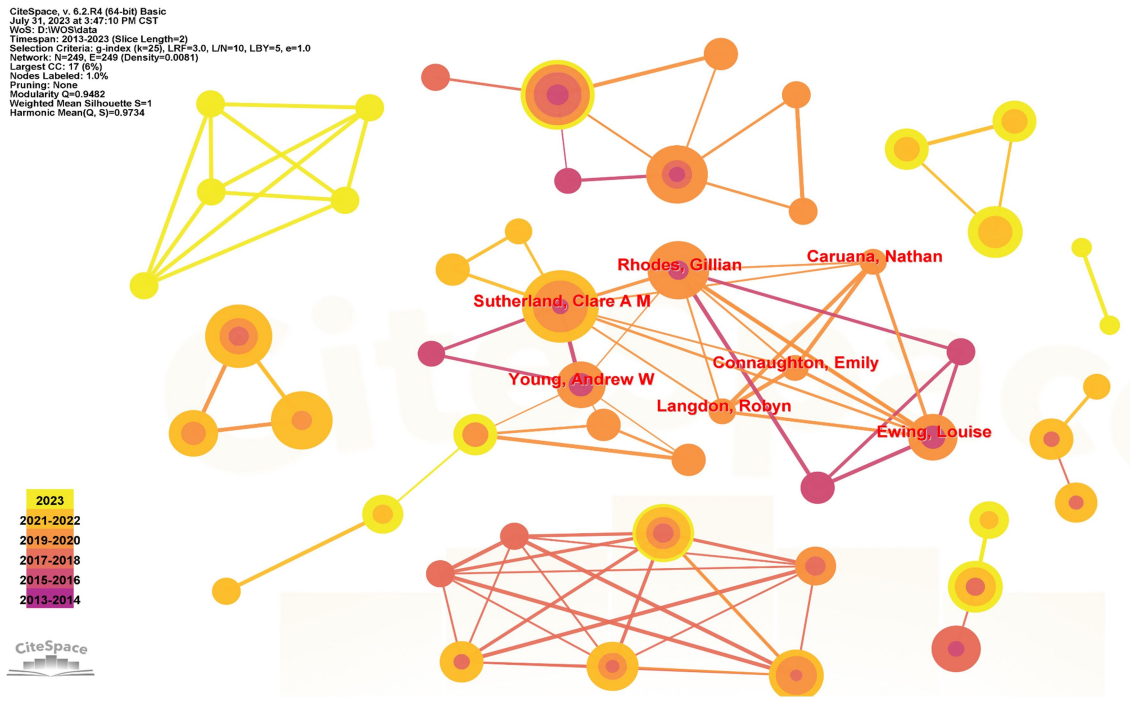


FIGURE 5
Author cooperation network diagram.

TABLE 3 Authors distribution.

Authors	Count	Authors	Burst	Years (2013–2023)
Todorov A	15	Brambilla M	3	
Sutherland CAM	14	Tipper SP	2.41	
Evans AM	12	Rhodes G	2.07	
Dotsch R	10	Jaeger B	2.03	
Rhodes G	9	Mieth L	2.03	
Jaeger B	9	Buchner A	2.03	
Alarcon GM	9	Topolinski S	1.98	
Brambilla M	8	Masi M	1.86	
Over H	8	Mattavelli S	1.86	
Lyons JB	8	Rule NO	1.81	

TABLE 4 Distribution of cited authors.

Authors	Count	Authors	Burst	Years (2013–2023)
Todorov A	415	Jaeger B	12.39	
Oosterhof NN	279	Ma DS	10.22	
Willis J	234	Glaeser EL	8.96	
Mayer RC	214	Debruine LM	8.20	
Zebrowitz LA	198	Eckel CC	7.24	
Fiske ST	185	Caulfield F	6.85	
Rule NO	161	Delgado MR	6.58	
Faul F	157	Simpson JA	6.56	
Berg J	153	Bayliss AP	6.28	
Colquitt JA	139	Suzuki A	6.25	

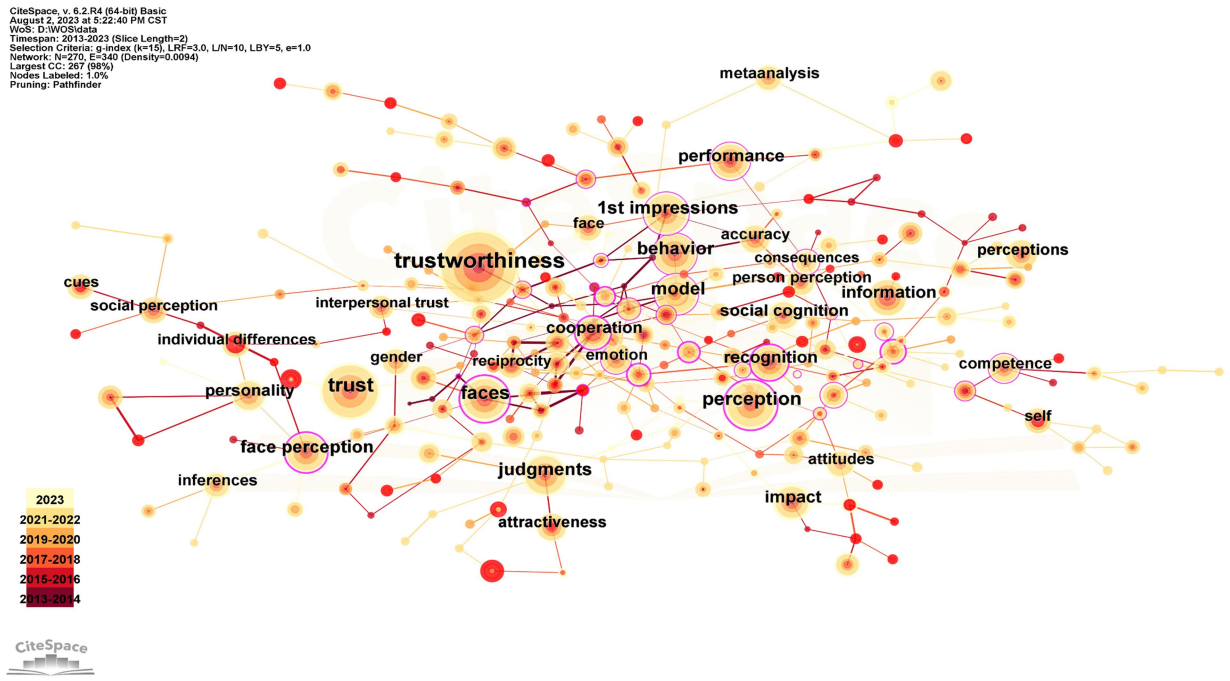


FIGURE 6
Keyword network diagram.

TABLE 5 Keyword distribution.

Keywords	Count	Keywords	Centrality
Trustworthiness	387	Competence	0.11
Trust	206	Model	0.10
Perception	161	Facial expressions	0.09
Faces	146	Faces	0.08
Judgments	125	Inferences	0.08
Behavior	123	Consequences	0.08
Model	122	Perspective	0.08
1st impressions	118	Children	0.08
Face perception	108	Age	0.08
Performance	93	Emotion	0.07

3.3.2 Research frontier analysis

Burst words are words that change significantly in quoted frequency in a certain period. Through the analysis of hot words, they can reflect the hotspots and frontier dynamics in a certain research field. A list of burst words was generated for the timeline by using CiteSpace and selecting the top 25 burst words of the study, as shown in Figure 7.

Until 2016, reputation, spontaneous trait inferences, and systems had attracted more and more attention. At this time, trustworthiness research was driven by reputation systems (e.g., Kuwabara, 2015; Wibral, 2015; Pouryazdan et al., 2016; Wang et al., 2016) as well as spontaneity (e.g., Klapper et al., 2016). By the time 2018, these several key words get more attention, like amygdala, cues, commitment, word of mouth, thin slices, self, organizational justice and men. Furthermore, the trustworthiness of research at this time concentrated on connections between tactics like emotional (e.g., Caulfield et al., 2015), customary culture (e.g., Sofer et al., 2017), and commitment (e.g., Kam et al., 2016). By 2020, the focus of trustworthiness research is mainly on experiences, mechanisms, individual differences, testimony, and quality. For example, how the audience's experience of sighting video in TV news affects the trustworthiness of reports (e.g., Halfmann et al., 2019); trust behavior and brain neurons (e.g., Wang et al., 2018; Zebrowitz et al., 2018). As of 2023, the current research focus has changed to health, determinants, race, time, culture, automation, acceptance, distrust, and scale. Trustworthiness research is not limited to the field of social communication, such as organizations, teachers, and students, but has gradually expanded to the medical field, for instance, health care, medical intelligence (Markus et al., 2021), and differences in trustworthiness between specific cultures or across cultures.

3.3.3 Research topic analysis

The cluster analysis of keywords based on the keyword distribution network is shown in Figure 5, further reveals the topic of trustworthiness research. In Figure 8, the cluster modularization Q value is 0.7778 ($Q > 0.30$), indicating a substantial cluster network association structure. In addition, the average contour value (S) is 0.9035, indicating that the cluster results are real and may act as a trustworthy source of data for trustworthiness studies. Overall fairness (marks of 0), artificial face (marks of 1), building trust (marks of 2), and unique clustering information constitute the 10 keyword clusters that emerged (see Table 6). After summarizing and combining the research hot spots in this field using clustering graph and clustering label related indicators, it is discovered that the research hot spots exhibit "multiple diffusion," which can be broadly classified into three core topics: facial cues, artificial intelligence, and

Top 25 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2013 - 2023
reputation	2013	3.31	2013	2016	
amygdala	2013	3.06	2013	2018	
cues	2015	3.73	2015	2018	
commitment	2015	3.5	2015	2018	
word of mouth	2016	3.36	2016	2018	
spontaneous trait inferences	2015	3.24	2015	2016	
systems	2015	3.17	2015	2016	
thin slices	2015	3.14	2015	2018	
self	2015	3.86	2017	2018	
experiences	2017	3.69	2017	2020	
organizational justice	2017	3.52	2017	2018	
mechanisms	2017	3.38	2017	2020	
men	2017	3.02	2017	2018	
individual differences	2015	4.76	2019	2020	
testimony	2019	4.06	2019	2020	
health	2017	3.72	2019	2023	
quality	2019	3.15	2019	2020	
determinants	2021	3.85	2021	2023	
race	2019	3.53	2021	2023	
time	2021	3.47	2021	2023	
culture	2021	3.47	2021	2023	
automation	2021	3.21	2021	2023	
acceptance	2021	3.08	2021	2023	
distrust	2021	3.08	2021	2023	
scale	2019	3.06	2021	2023	

FIGURE 7
Keywords with the strongest citation bursts.

social perception. This will be covered in depth in the discussion section.

4 Discussion

Regarding positive expectations of the other party's intentions and behavior, the extent to which one party is willing to take risks or expose itself to vulnerabilities is termed trust (Mayer et al., 1995). Trust is closely attached to interpersonal interactions such as reciprocity, cooperation, and betrayal (Lemmers-Jansen et al., 2017), and it serves an essential and dominant role in individual behavior (Falk and Hermle, 2018). People choose whether to act on trust based on perceived legitimacy; thus, trust does not just emerge out of thin air. Breuer and McDermott (2010) argue that trustworthiness is more important than trust for the success of public policy and sustainable long-term economic growth. In part because trustworthiness supports trust.

Most of the quantitative research on trustworthiness focus on a certain field, and our analysis has a basic understanding of the general framework of trustworthiness research through the citation knowledge graph. Most of the quantitative research on trustworthiness focus on a certain field, and our analysis has a basic understanding of the general framework of trustworthiness research through the citation knowledge graph. The results of this study show that the number of trustworthiness related studies has increased

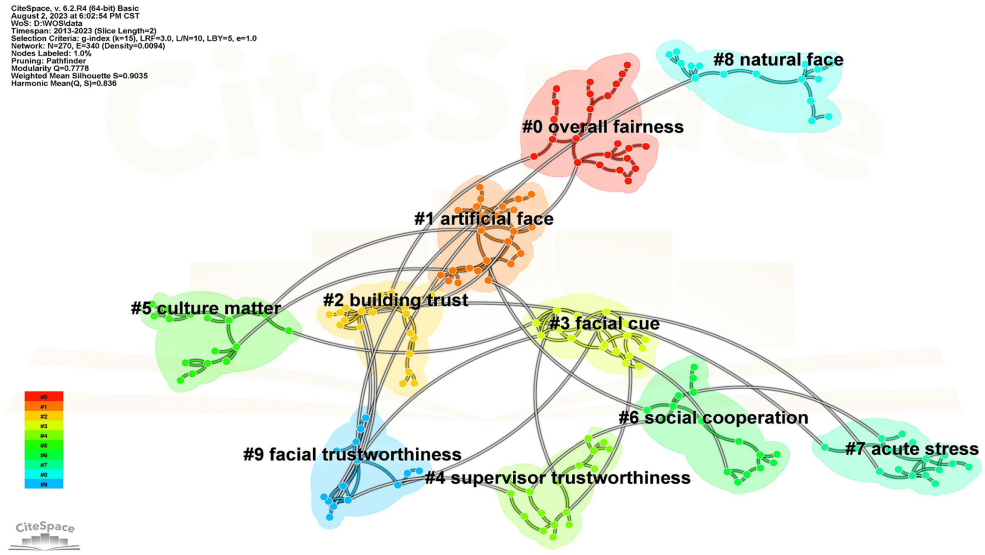


FIGURE 8
Keyword clustering network diagram.

TABLE 6 Keyword clustering information.

Cluster ID	Size	Silhouette	Mean (Year)	(Label) LLR
0	24	0.951	2017	Qualitative research; trust; artificial intelligence; justice; procedural justice
1	24	0.926	2017	First impressions; emotional expressions; facial trustworthiness; face recognition; artificial faces
2	20	0.899	2017	Face perception; risk; person perception; social preferences; trust
3	19	0.951	2015	Facial expression; emotion; smile; gaze cueing; happiness
4	17	0.915	2017	Person perception; social cognition; games; trait inferences; distrust
5	17	0.916	2018	Competence; warmth; social competence; cultural differences; stereotype
6	16	0.937	2016	Facial attractiveness; attractiveness; makeup; evolutionary psychology; physical attractiveness
7	16	0.796	2017	Psychology; others; economic games; halo effect; youth
8	16	0.952	2017	Face perception; social perception; open data; emotion recognition; open materials
9	15	0.901	2015	Evolution; expressions; thin slices; schizophrenia; continuous flash suppression

generally in the past decade; The trustworthiness research mainly focuses on the industrialized Europe and the United States, in which the research results of the United States have greater global influence; The University of California system, Harvard and Yale are among the most prolific institutions; The core authors are outstanding university scholars, represented by Alexander Todorov and others, but the level of cooperation among the core authors needs to be improved. The main journals that have published trustworthiness studies are *the Journal of Personality and Social Psychology* and *Biology Letters*. This report shows that cutting-edge research can be employed to divide trustworthiness-related research into three research directions: facial clues, artificial intelligence, and social perception. ABI model theory is a relatively popular and foundational theory for understanding trustworthiness, and although it was initially rooted in the context of trust within organizations, researchers have applied this model to a range of contexts, and there are many positive correlations between the three factors, so the ABI model is closely related to trustworthiness related research topics.

4.1 Facial cues and trustworthiness

Among the many factors that affect trustworthiness, facial clues have always been a hot topic of concern to researchers. Since the start of the 20th century, psychologists have known that there is general agreement that facial features are related to social and personality traits (Todorov et al., 2015). Face typicality is an important factor in social perception because it influences trustworthiness judgments. And the trustworthiness judgment is like the basic evaluation of the human face (Sofer et al., 2015). Wilson and Rule (2015) demonstrated how perceptions of people's faces might be biased and influence their daily lives. According to Rhodes et al. (2012), facial indications have an early impact on trust behavior, and 10-year-olds preferentially trust partners they perceive to be trustworthy. The findings of Li et al. (2023) suggest that when one learns that another person is trustworthy (or unbelievable), the corresponding graphic traits in the mind are overlaid on the physical characteristics of the individual's face. Then the facial characteristics are reshaped.

The current study also explores whether the perception of another person's trustworthiness affects the characterization of the other person's facial appearance and its potential mechanisms. At the same time, COVID-19 has made wearing masks common. When judging attractiveness, masks can enhance the attractiveness of less beautiful faces, but can reduce the attractiveness of more beautiful faces (Wang et al., 2023). Although one can form a stable first impression based on facial and vocal cues, their accuracy is low. Voice-based first impressions tend to be more positive than face-based first impressions (Jiang et al., 2024). In the study of facial trustworthiness, researchers have accumulated many theories (such as typical emotional generalization theory and typical theory) and experience. Future research on facial trustworthiness may be even deeper.

4.2 Artificial intelligence and trustworthiness

Intelligent technologies are increasingly entering the workplace. It has gradually shifted from workflow-supporting technologies to artificial intelligence (AI) agents as team members. And it has great potential in improving the health and well-being of the people (Ulfert et al., 2023). Although there are few applications for robots in clinical practice, they can benefit older people by reducing loneliness, troublesome behavior, and depression and improving social contact (Broadbent, 2017). Markus et al. (2021) discovered a lack of transparency as one of the major barriers to the clinical application of AI. They argue that explainable modeling can support reliable AI, though there was still an absence of useful evidence. But it may be used to support additional steps, such as reporting data quality, implementing extensive (external) validation, and regulation, to create trustworthy artificial intelligence. In the opinion of Song et al. (2023), regulatory compatibility expanded throughout par-social interaction and was a key element in the activation of social robot trustworthiness. To address the cognitive and emotional demands of users, artificial intelligence can also be used in the field of psychotherapy. It can simulate a variety of mental talents, including not just advanced processing and memory but also a few basic social and emotional abilities (Wiese et al., 2022).

Today, chatbot technology is constantly changing the interactive experience of traditional unguided online therapeutic intervention programs. It provides both human-like guidance and achieves full automation (Mo et al., 2023). However, researchers know very little about why chatbots operate, and there are currently no researchers to compile real, effective relationship clues to guide the design of chatbots. As a result, the investigation into the trustworthiness of AI may be in accordance with the evolving trends of new AI technologies.

4.3 Social perception and trustworthiness

The evaluation of trustworthiness of others included three aspects: ability, integrity, and benevolence, which could affect the perception of trustworthiness. When faced with integrity-benevolence and moral conflict, the individual's trust behavior was also affected. For instance, Lupoli et al. (2020) demonstrated that while being viewed as having compassion can be a sign of goodwill, it does not always foster trust when presented with a moral dilemma. In addition, differences in trust and trustworthiness between cultures fell within the study. The study by Huang and Rau (2019) examined the impact of trust and trust in cooperation with friends or strangers in two different cultural business environments. The results revealed that Chinese and American participants had higher levels of trust and trustworthiness in friends than strangers. And Chinese participants were better able to distinguish between friends and strangers than American participants.

The degree of trust in both people and institutions is influenced by trustworthiness. Recent years have seen an increase in the frequency of emergencies, and how an organization responds to social emergencies has a bearing on its trustworthiness and the public's level of trust in it. Emergencies are typically connected to institutions like governments. When people blame the government for environmental problems, their trust in the government declines (Kentmen, 2013). When negative events are not officially or authoritative, people are more faith in conspiracy theories (Xie et al., 2022). The result of liability attribution also has an impact on its relationship with the people (Ma and Zhan, 2016). Such research might strengthen the pillars of trustworthiness-related research further, opening the door for trustworthy applied research.

4.4 Analysis of the above three topics based on ABI model

Mayer and his colleagues conceptualize the trustworthiness structure as three interconnected factors: ability, benevolence, and integrity, which together determine whether a person or organization is trustworthy. Perceived ability is defined as the belief that a fiduciary can perform one or more specific tasks. Perceived benevolence is defined as the trustee's perceived willingness to act in the best interests of the principal. Perceived integrity is defined as the degree to which the trustee's values are believed to be compatible with their own. Mayer et al. (1995) present a complete theoretical framework for the concept of trustworthiness. Artificial intelligence, facial expressions, and social perception are all explainable using the ABI theoretical model.

Based on the ABI model's trustworthiness and AI, Trust is a critical necessity for efficient human-computer interaction, as artificial organisms integrate into human civilization in a social setting. To fully integrate into our culture and optimize their acceptance and trustworthiness, artificial agents must adapt to the intricacies of their surroundings, just as people do. In a study of human-AI collaboration, indications of ability, warmth, and integrity influenced trustworthiness (Jorge et al., 2024). The use of artificial intelligence in psychotherapy necessitates replicating a wide range of psychological skills, including not only superior processing and memory but also some fundamental social and emotional capacities (Wiese et al., 2022). One of the key challenges to clinical AI application is a lack of openness in terms of data quality reporting, thorough (external) validation, and regulation to build trusted AI (Markus et al., 2021). Having advanced processing performance, social and emotional capabilities, and external oversight increases transparency, which corresponds to the three characteristics of trustworthiness.

Based on the ABI model's trustworthiness and facial clues, many visual indicators, including facial expression and gender, influence

people's trustworthiness judgments at the same time. Entrepreneurs' facial trustworthiness is positively correlated with the success of crowdfunding campaigns (Duan et al., 2020), and when a person's facial expression conveys confidence and professionalism, others are more likely to believe that the person possesses the necessary skills and knowledge to complete the task. Happiness enhances the impression of trustworthiness, whereas anger diminishes it (Oosterhof and Todorov, 2009), and warm smiles, eye contact, and sympathetic expressions can also serve as indications of friendliness (e.g., Li et al., 2021). Angry facial expressions indicate immediate potential threats, and adults may predict violence and aggression based on face structure (Short et al., 2012).

Based on the ABI model's trustworthiness and social perception. Ability, benevolence, and integrity influence trustworthiness in both individuals and organizations; yet, the definitions of trust and trustworthiness are sometimes implicit, and it may be unclear who or what is trusted. When trustees' social identities increased, they were deemed more trustworthy (Xin and Zhang, 2018); prosocial liars are sometimes perceived as more trustworthy (Levine and Schweitzer, 2014).

5 Limitations and future research

5.1 Limitations

Although we conducted a topic search, so that the papers examined are the most relevant, The main drawback of co-citation analyses is the impossibility of fully collecting and displaying the entire existing literature (Stehmann, 2020). Firstly, only CiteSpace, a measurement analysis tool, and other readily available databases (Scopus, PubMed, etc.) and analytical tools (such as VOSviewer). Secondly, literature filtering duration is only 10 years and does not cover all relevant literature, were utilized in this study's literature analysis of just one Web of Science database. Thirdly, the study is primarily based on empirical research and only from the realm of psychology, may not fully represent qualitative or transdisciplinary perspectives on trustworthiness. And lastly, the cited references list only the first authors instead of all authors, the citation rate does not reflect contributions of the second or further authors, which could affect citations accuracy regarding some authors (Garfield, 1979).

5.2 Future research

Since brain imaging technology has advanced, researchers have focused on the cognitive neural mechanisms underlying trustworthiness. They have discovered that, in addition to the almond nucleus, other brain regions-such as the internal frontal cortex and the right hip joint region—are also active during trust decision-making (Euston et al., 2023). Bellucci et al. (2019) combined a new paradigm for successfully inducing impressions of confidence through functional MRI and multivariate analysis. Studies have demonstrated integrity-based trustworthiness performance in the posterior cingulate cortex, dorsolateral prefrontal cortex, and intraparietal sulcus. Brain signals in these regions can predict the individual's trust in subsequent social interactions with the same partner. Sijtsma et al. (2023) used functional magnetic resonance imaging (FMRI) to provide insights into how the behaviors and neural mechanisms of adolescent trust are affected by expectations. In the meantime, studies by Frazier et al. (2021) reinforced the idea that aging lessens sensitivity to traces of trustworthiness, but that intranasal oxytocin has no effect on behavioral adjustment.

In addition to FMRI techniques, event-related potential techniques (ERPs) have been further explored by using neural indicators reflecting electrophysiological activity in the cerebral cortex. P1, N17, early post-negative voltage (EPN), late positive component (LPC), and feedback negative waves (FN) have now been found to be important ERP indicators in this field (Leng et al., 2020). These findings grow our understanding of the neurological underpinnings of specific social traits. However, the accuracy and ecological usefulness of the results are negatively impacted by the measurement patterns and experimental materials, which are more uniform and primarily consist of static faces. Additionally, it is not clear which specific perceptual information—such as emotional cues or typicity—contributes more to brain region activation. With the development and combination of various technical means, this field can be discussed with more scientific and rigorous methods. In the future, researchers can continue to study the mechanism of the influence of trustworthiness on trust decisions through dynamic faces, to deeply explore the mechanism of the influence of trustworthiness on trust decisions.

In the case of trustworthiness-related empirical studies, it is typically divided between the trustor (the party whose trust has been violated) and the trustee (the one who carries out a trust violation). The relationship is clear and trust relationship is just one form of interpersonal relationships. However, the boundary between the responsibility subject and the responsibility is not so clear, especially in the collectivist environment like China. The situation may be more complex, and whether the research results can be extended to the real situation is worth further investigation.

Researchers have produced many experiences and results on trustworthiness studies. It is the focus of the field of economics and organization, and gradually radiated to the field of education. Brodsky et al. (2021) find that fact-checking strategies for improving college students through lateral reading teaching in general education civic courses require further research. Future research is needed to determine whether the improvement in lateral reading is maintained over time and to explore other factors. List and Oaxaca (2023) try to ascertain the effectiveness of college students participating in study report critique, examining the effects and potential future directions of student critical capacity development. The study of Su and Kong (2023) find that Chinese music students in English will encounter some severe challenges in the teaching language course (EMI) due to their limited English proficiency. Alves-Wold et al. (2023) studied the writing motivation of K-5 students. However, because of the lack of teacher behavioral perspectives and the main emphasis on higher education in these studies, future research may take a more important turn when combined with the findings of the hot spots and effective keyword analysis. For instance, the trustworthiness of educators with various cultural backgrounds and grade levels is studied under the update of education policy, and how to apply it in practice is considered, such as the update of the evaluation system and steps to improve trustworthiness.

6 Conclusion

This study demonstrates that the number of trustworthiness related studies has increased generally in the past decade; The University of California system, Harvard and Yale are among the most prolific institutions; The core authors represented by Alexander Todorov and others. The main journals that have published trustworthiness studies are *the Journal of Personality and Social Psychology* and *Biology Letters*. Popular topics include facial trustworthiness, brain neurology, medical trustworthiness, and cultural differences. Three factors inform the hot spot direction: facial clues, artificial intelligence, and social perception.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

ZZ: Funding acquisition, Methodology, Writing – review & editing. WD: Formal analysis, Methodology, Writing – original draft. YW: Formal analysis, Methodology, Writing – original draft. CQ: Conceptualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (32000754), the Youth Foundation of the Ministry of Education of Humanities and Social Science Project of China (20YJC190030), and the Henan Province Higher Education Youth Backbone Teacher Training Project (2023GGJS039).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alves-Wold, A., Walgermo, B. R., McTigue, E., and Uppstad, P. H. (2023). Assessing writing motivation: a systematic review of K-5 Students' self-reports. *Educ. Psychol. Rev.* 35:24. doi: 10.1007/s10648-023-09732-6

Bai, C., Gong, Y., and Feng, C. (2019). Social trust, pattern of difference, and subjective well–being. *SAGE Open* 9:215824401986576. doi: 10.1177/2158244019865765

Bailey, P. E., and Leon, T. (2019). A systematic review and meta-analysis of age-related differences in trust. *Psychol. Aging* 34, 674–685. doi: 10.1037/pag0000368

Bellucci, G., Molter, F., and Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. *Nat. Commun.* 10:5184. doi: 10.1038/s41467-019-13261-8

Bennett, M. (2023). Trusting groups. *Philos. Psychol.* 37, 196–215. doi: 10.1080/09515089.2023.2179478

Best, A. L., Fletcher, F. E., Kadono, M., and Warren, R. C. (2021). Institutional distrust among African Americans and building trustworthiness in the COVID-19 response: implications for ethical public health practice. *J. Health Care Poor Underserved* 32, 90–98. doi: 10.1353/hpu.2021.0010

Bicchieri, C., Xiao, E., and Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Polit. Philos. Econ.* 10, 170–187. doi: 10.1177/1470594x10387260

Bottom, W. P., Gibson, K., Daniels, S. E., and Murighan, J. K. (2002). When talk is not cheap: substantive penance and expressions of intent in rebuilding cooperation. *Organ. Sci.* 13, 497–513. doi: 10.1287/orsc.13.5.497.7816

Breuer, J. B., and McDermott, J. (2010). Trustworthiness and economic performance. SSRN [Preprint].

Broadbent, E. (2017). Interactions with robots: the truths we reveal about ourselves. *Annu. Rev. Psychol.* 68, 627–652. doi: 10.1146/annurev-psych-010416-043958

Brodsky, J. E., Brooks, P. J., Scimeca, D., Todorova, R., Galati, P., Batson, M., et al. (2021). Improving college students' fact–checking strategies through lateral reading instruction in a general education civics course. *Cog. Res. Princ. Imp.* 6, 1–18. doi: 10.1186/s41235-021-00291-4

Caulfield, F., Ewing, L.Bank, S, and Rhodes, G. (2015). Judging trustworthiness from faces: emotion cues modulate trustworthiness judgments in young children. *Br. J. Psychol.* 107, 503–518. doi: 10.1111/bjop.12156

Cheer, K., MacLaren, D., and Tsey, K. (2015). The use of grounded theory in studies of nurses and midwives' coping processes: a systematic literature search. *Contemp. Nurse* 51, 200–219. doi: 10.1080/10376178.2016.1157445

Cottrell, C. A., Neuberg, S. L., and Li, N. P. (2007). What do people desire in others? A socio functional perspective on the importance of different valued characteristics. *J. Pers. Soc. Psychol.* 92, 208–231. doi: 10.1037/0022-3514.92.2.208

Deutsch, M. (1958). Trust and suspicion. *J. Confl. Resolut.* 2, 265–279. doi: 10.1177/002200275800200401

Du, P. C., Nguyen, M. H. B., Foulk, T. A., and Schaerer, M. (2023). Relative power and interpersonal trust. *J. Pers. Soc. Psychol.* 124, 567–592. doi: 10.1037/pspi0000401

Duan, Y., Hsieh, T. S., Wang, R. R., and Wang, Z. (2020). Entrepreneurs' facial trustworthiness, gender, and crowdfunding success. *J. Corp. Financ.* 64:101693. doi: 10.1016/j.jcorpfin.2020.101693

Euston, D. R., Gruber, A. J., and McNaughton, B. L. (2023). The role of medial prefrontal cortex in memory and decision making. *Neuron* 76, 1057–1070. doi: 10.1016/j.neuron.2012.12.002

Falk, A., and Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science* 362:eaas9899. doi: 10.1126/science.aas9899

Filieri, R. (2016). What makes an online consumer review trustworthy? *Ann. Tour. Res.* 58, 46–64. doi: 10.1016/j.annals.2015.12.019

Frazier, I., Lin, T., Liu, P., Skarsten, S., Feifel, D., and Ebner, N. C. (2021). Age and intranasal oxytocin effects on trust–related decisions after breach of trust: behavioral and brain evidence. *Psychol. Aging* 36, 10–21. doi: 10.1037/pag0000545

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics* 1, 359–375. doi: 10.1007/BF02019306

Halfmann, A., Dech, H., Riemann, J., Schlenker, L., and Wessler, H. (2019). Moving closer to the action: how viewers' experiences of eyewitness videos in TV news influence the trustworthiness of the reports. *J. Mass Commun. Q.* 96, 367–384. doi: 10.1177/1077699018785890

Huang, H., and Rau, P. L. P. (2019). Cooperative trust and trustworthiness in China and the United States: does guanxi make a difference? *Soc. Behav. Pers.* 47, 1–11. doi: 10.2224/sbp.7779

Jia, G. L., Ma, R. G., and Hu, Z. H. (2019). Review of urban transportation network design problems based on CiteSpace. *Math. Probl.* 2019, 1–22. doi: 10.1155/2019/5735702

Jiang, Z., Li, D., Li, Z., Yang, Y., Liu, Y., Yue, X., et al. (2024). Comparison of face-based and voice-based first impressions in a Chinese sample. *Br. J. Psychol.* 115, 20–39. doi: 10.1111/bjop.12675

Jorge, C. C., Jonker, C. M., and Tielman, M. L. (2024). How should an AI trust its human teammates? Exploring possible cues of artificial trust. *ACM Trans Interact Intell Syst.* 14, 1–26. doi: 10.1145/3635475

Kam, C., Morin, A. J., Meyer, J. P., and Topolnytsky, L. (2016). Are commitment profiles stable and predictable? A latent transition analysis. *J. Manage.* 42, 1462–1490. doi: 10.1177/0149206313503010

Kennedy, J. A., and Schweitzer, M. E. (2018). Building trust by tearing others down: when accusing others of unethical behavior engenders trust. *Organ. Behav. Hum. Decis. Process.* 149, 111–128. doi: 10.1016/j.obhdp.2018.10.001

Kentmen, C. C. (2013). Blaming the government for environmental problems: a multilevel and cross–national analysis of the relationship between trust in government and local and global environmental concerns. *Environ. Behav.* 45, 971–992. doi: 10.1177/0013916512453840

Kim, P. H., Dirks, K. T., and Cooper, C. D. (2009). The repair of trust: a dynamic bilateral perspective and multilevel conceptualization. *Acad. Manage. Rev.* 34, 401–422. doi: 10.5465/amr.2009.40631887

Kim, P. H., Ferrin, D. L., Cooper, C. D., and Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence– versus integrity–based trust violations. *J. Appl. Psychol.* 89, 104–118. doi: 10.1037/0021-9010.89.1.104

Klapper, A., Dotsch, R., van Rooij, I., and Wigboldus, D. H. J. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *J. Pers. Soc. Psychol.* 111, 655–664. doi: 10.1037/pspa0000062

Kuwabara, K. (2015). Do reputation systems undermine trust? Divergent effects of enforcement type on generalized trust and trustworthiness. *Am. J. Sociol.* 120, 1390–1428. doi: 10.1086/681231

Lemmers-Jansen, I. L. J., Krabbendam, L., Veltman, D. J., and Fett, A. K. J. (2017). Boys vs. girls: gender differences in the neural development of trust and reciprocity depend on social context. *Dev. Cogn. Neurosci.* 25, 235–245. doi: 10.1016/j.dcn.2017.02.001

Leng, H., Liu, Y., Li, Q., Wu, Q., and Li, D. (2020). Outcome evaluation affects facial trustworthiness: an event-related potential study. *Front. Hum. Neurosci.* 14:514142. doi: 10.3389/fnhum.2020.514142

Levine, E. E., Bitterly, T. B., Cohen, T. R., and Schweitzer, M. E. (2018). Who is trustworthy? Predicting trustworthy intentions and behavior. *J. Pers. Soc. Psychol.* 115, 468–494. doi: 10.1037/pspi0000136

Levine, E. E., and Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *J. Exp. Soc. Psychol.* 53, 107–117. doi: 10.1016/j.jesp.2014.03.005

Li, Q., Fang, W., Hu, C., Shi, D., Hu, X., Fu, G., et al. (2023). Can Cinderella become snow white? The influence of perceived trustworthiness on the mental representation of faces. *Acta Psychol. Sin.* 55, 1518–1528. doi: 10.3724/SP.J.1041.2023.01518

Li, Y., Jiao, X., Liu, Y., Tse, C. S., and Dong, Y. (2021). Age differences in facial trustworthiness judgement based on multiple facial cues. *Br. J. Psychol.* 112, 474–492. doi: 10.1111/bjop.12472

List, A., and Oaxaca, G. S. C. (2023). Comprehension and critique: an examination of students' evaluations of information in texts. *Read. Writ.* 37, 641–671. doi: 10.1007/s11145-023-10417-3

Liu, Z., Wang, X., and Chen, C. (2009). Scientific knowledge graph method and its application in scientific and technological information. *Digit. Library. Forum.* 10, 14–34. doi: 10.3772/j.issn.1673-2286.2009.10.004

Lleó de Nalda, A., Guillen, M., and Gil Pechuan, I. (2016). The influence of ability, benevolence, and integrity in trust between managers and subordinates: the role of ethical reasoning. *Bus. Ethics. Eur. Rev.* 25, 556–576. doi: 10.1111/beer.12117

Lupoli, M. J., Zhang, M., Yin, Y., and Oveis, C. (2020). A conflict of values: when perceived compassion decreases trust. *J. Exp. Soc. Psychol.* 91:104049. doi: 10.1016/j.jesp.2020.104049

Ma, L., and Zhan, M. (2016). Effects of attributed responsibility and response strategies on organizational reputation: a meta–analysis of situational crisis communication theory research. *J. Public Relat. Res.* 28, 102–119. doi: 10.1080/1062726X.2016.1166367

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* 113:103655. doi: 10.1016/j.jbi.2020.103655

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 709–734. doi: 10.2307/258792

Milesi, A., De Carli, P., Locati, F., Benzi, I., Campbell, C., Fonagy, P., et al. (2023). How can I trust you? The role of facial trustworthiness in the development of epistemic and interpersonal trust. *Hum. Dev.* 67, 57–68. doi: 10.1159/000530248

Mo, R., Fang, Z., and Fang, J. (2023). How to establish a digital therapeutic alliance between chatbots and users: the role of relational cues. *Adv. Psychol. Sci.* 31, 669–683. doi: 10.3724/SP.J.1042.2023.00669

Oosterhof, N. N., and Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion* 9, 128–133. doi: 10.1037/a0014520

Poon, J. M. (2013). Effects of benevolence, integrity, and ability on trust-in-supervisor. *Empl. Relat.* 35, 396–407. doi: 10.1108/er-03-2012-0025

Pouryazdan, M., Kantarci, B., Soyata, T., and Song, H. (2016). Anchor–assisted and vote–based trustworthiness assurance in smart city crowdsensing. *IEEE. Access.* 4, 529–541. doi: 10.1109/ACCESS.2016.2519820

Qiu, J., Duan, Y., Chen, J., Song, E., and Ji, L. (2003). The retrospect and prospect on bibliometrics in China. *J. Sci. Res.* 21, 143–148. doi: 10.16192/j.cnki.1003-2053.2003.02.007

Radke, S., Kalt, T., Wagels, L., and Derntl, B. (2018). Implicit and explicit motivational tendencies to faces varying in trustworthiness and dominance in men. *Front. Behav. Neurosci.* 12:8. doi: 10.3389/fnbeh.2018.00008

Reimann, M., Hüller, C., Schilke, O., and Cook, K. S. (2022). Impression management attenuates the effect of ability on trust in economic exchange. *Proc. Natl. Acad. Sci. U. S. A.* 119:e2118548119. doi: 10.1073/pnas.2118548119

Rhodes, G., Morley, G., and Simmons, L. W. (2012). Women can judge sexual unfaithfulness from unfamiliar men's faces. *Biol. Lett.* 9:20120908. doi: 10.1098/rsbl.2012.0908

Santos, S., Almeida, I., Oliveiros, B., and Castelo-Branco, M. (2016). The role of the amygdala in facial trustworthiness processing: a systematic review and meta-analyses of fMRI studies. *PLoS One* 11:e0167276. doi: 10.1371/journal.pone.0167276

Ścigała, K. A., Schild, C., and Zettler, I. (2020). Dishonesty as a signal of trustworthiness:honesty–humility and trustworthy dishonesty. *R. Soc. Open Sci.* 7:200685. doi: 10.1098/rsos.200685

Shayo, H. J., Rao, C., and Kakupa, P. (2021). Conceptualization and measurement of trust in home–school contexts: a scoping review. *Front. Psychol.* 12:742917. doi: 10.3389/fpsyg.2021.742917

Short, L. A., Mondloch, C. J., McCormick, C. M., Carré, J. M., Ma, R., Fu, G., et al. (2012). Detection of propensity for aggression based on facial structure irrespective of face race. *Evol. Hum. Behav.* 33, 121–129. doi: 10.1016/j.evolhumbehav.2011.07.002

Shu, G., Meng, S., and Xiao, N. (2021). Impact of trust violations on attentional Bias and working memory updating. *Curr. Psychol.* 42, 967–979. doi: 10.1007/s12144-021-01432-0

Siddique, S., Sutherland, C. A., Palermo, R., Foo, Y. Z., Swe, D. C., and Jeffery, L. (2022). Development of face-based trustworthiness impressions in childhood: a systematic review and metaanalysis. *Cogn. Dev.* 61:101131. doi: 10.1016/j.cogdev.2021.101131

Sijtsma, H., Lee, N. C., van Kesteren, M. T. R., Braams, B. R., van Atteveldt, N. M., Krabbendam, L., et al. (2023). The effect of incorrect prior information on trust behavior in adolescents. *Neuropsychologia* 179:108423. doi: 10.1016/j.neuropsychologia.2022.108423

Siuda, S., Schlösser, T., and Fetchenhauer, D. (2022). Do we know whom to trust? A review on trustworthiness detection accuracy. *Int. Rev. Soc. Psychol.* 35, 1–16. doi: 10.5334/irsp.623

Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H., and Todorov, A. (2017). For your local eyes only: culture–specific face typicality influences perceptions of trustworthiness. *Perception* 46, 914–928. doi: 10.1177/0301006617691786

Sofer, C., Dotsch, R., Wigboldus, D. H. J., and Todorov, A. (2015). What is typical is good: the influence of face typicality on perceived trustworthiness. *Psychol. Sci.* 26, 39–47. doi: 10.1177/0956797614554955

Song, Y., and Luximon, Y. (2020). Trust in AI agent: a systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors* 20:5087. doi: 10.3390/s20185087

Song, Y., Tao, D., and Luximon, Y. (2023). In robot we trust? The effect of emotional expressions and contextual cues on anthropomorphic trustworthiness. *Appl. Ergon.* 109:103967. doi: 10.1016/j.apergo.2023.103967

Stehmann, J. (2020). Identifying research streams in online gambling and gaming literature: a bibliometric analysis. *Comput. Hum. Behav.* 107:106219. doi: 10.1016/j.chb.2019.106219

Su, P., and Kong, J. (2023). Implementing EMI in Chinese music classes: students' perceived benefits and challenges. *Front. Psychol.* 14:1086392. doi: 10.3389/fpsyg.2023.1086392

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., and Young, A. W. (2013). Social inferences from faces: ambient images generate a three–dimensional model. *Cognition* 127, 105–118. doi: 10.1016/j.cognition.2012.12.001

Todorov, A., Olivola, C. Y., Dotsch, R., and Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* 66, 519–545. doi: 10.1146/annurev-psych-113011-143831

Tomlinson, E. C., Schnackenberg, A. K., Dawley, D., and Ash, S. R. (2020). Revisiting the trustworthiness–trust relationship: exploring the differential predictors of cognitionand affect–based trust. *J. Organ. Behav.* 41, 535–550. doi: 10.1002/job.2448

Travers, M. J., Murphy, M. C., Debenham, J. R., Chivers, P., Bulsara, M. K., Bagg, M. K., et al. (2019). Should this systematic review and meta-analysis change my practice? Part 1: exploring treatment effect and trustworthiness. *Br. J. Sports Med.* 53, 1488–1492. doi: 10.1136/bjsports-2018-099958

Ulfert, A. S., Georganta, E., Centeio Jorge, C., Mehrotra, S., and Tielman, M. (2023). Shaping a multidisciplinary understanding of team trust in human–AI teams: a theoretical framework. *Eur. J. Work Organ. Psy.* 33, 158–171. doi: 10.1080/1359432X.2023.2200172

van 't Wout, M., and Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision–making. *Cognition* 108, 796–803. doi: 10.1016/j.cognition.2008.07.002

van der Werff, L., O'Shea, D., Healy, G., Buckley, F., Real, C., Keane, M., et al. (2023). The neuroscience of trust violation: differential activation of the default mode network in ability, benevolence and integrity breaches. *Appl. Psychol.* 72, 1392–1408. doi: 10.1111/apps.12437

Wang, S., Falvello, V., PorterJ Said, C. P., and Todorov, A. (2018). Behavioral and neural adaptation in approach behavior. *J. Cogn. Neurosci.* 30, 885–897. doi: 10.1162/jocn_a_01243

Wang, S., Han, C., Sang, Z., Zhang, X., Chen, S., Wang, H., et al. (2023). Hidden faces, altered perceptions: the impact of face masks on interpersonal perception. *Front. Psychol.* 14:1203442. doi: 10.3389/fpsyg.2023.1203442

Wang, S., Huang, L., Hsu, C. H., and Yang, F. (2016). Collaboration reputation for trustworthy web service selection in social networks. *J. Comput. Syst. Sci.* 82, 130–143. doi: 10.1016/j.jcss.2015.06.009

Wibral, M. (2015). Identity changes and the efficiency of reputation systems. *Exp. Econ.* 18, 408–431. doi: 10.1007/s10683-014-9410-3

Wiese, E., Weis, P. P., Bigman, Y., Kapsaskis, K., and Gray, K. (2022). It's a match: task assignment in human–robot collaboration depends on mind perception. *Int. J. Soc. Robot.* 14, 141–148. doi: 10.1007/s12369-021-00771-z

Wilson, J. P., and Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal–sentencing outcomes. *Psychol. Sci.* 26, 1325–1331. doi: 10.1177/0956797615590992

Xie, X., Zhang, Y., and Guo, Y. (2022). Psychological needs of responsibility attribution and response strategies in public emergencies. *Adv. Psychol. Sci.* 30, 1327–1335. doi: 10.3724/SP.J.1042.2022.01327

Xin, Z., and Zhang, Y. (2018). The impact of the number of a trustee's social identities on their trustworthiness. *J. Pac. Rim Psychol.* 12:e30. doi: 10.1017/prp.2018.15

Yan, Y., and Wu, X. (2016). From trust violation to trust repair: the role of moral emotions. *Adv. Psychol. Sci.* 24, 633–642. doi: 10.3724/SP.J.1042.2016.00633

Yang, Q., and Beatty, M. (2016). A meta-analytic review of health information credibility: belief in physicians or belief in peers? *Health Inf. Manag. J.* 45, 80–89. doi: 10.1177/1833358316639432

Zebrowitz, L. A., Ward, N., Boshyan, J., Gutchess, A., and Hadjikhani, N. (2018). Older adults' neural activation in the reward circuit is sensitive to face trustworthiness. *Cogn. Affect. Behav. Neurosci.* 18, 21–34. doi: 10.3758/s13415-017-0549-1

Zhu, N., Jiang, N., and Liu, Y. (2022). The development of employees' feeling trusted by their supervisors. *Adv. Psychol. Sci.* 30, 1448–1462. doi: 10.3724/SP.J.1042.2022.01448

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 03 January 2025
DOI 10.3389/fpsyg.2024.1440768

OPEN ACCESS

EDITED BY
Carmelo Mario Vicario,
University of Messina, Italy

REVIEWED BY
Shuailei Lian,
Yangtze University, China
Björn Sjögren,
Linköping University, Sweden

*CORRESPONDENCE
Chunhui Qi
✉ qchizz@126.com

RECEIVED 30 May 2024
ACCEPTED 13 December 2024
PUBLISHED 03 January 2025

CITATION
Zhang Z, Cai X, Gao W, Zhang Z and Qi C (2025) The impact of moral judgment on bystanders' interpersonal trust: the mediating role of trustworthiness.
Front. Psychol. 15:1440768.
doi: 10.3389/fpsyg.2024.1440768

The impact of moral judgment on bystanders' interpersonal trust: the mediating role of trustworthiness

Zhen Zhang[1,2], Xia Cai[1], Weiwei Gao[3], Zengtong Zhang[1] and Chunhui Qi[1]*

[1]Faculty of Education, Henan Normal University, Xinxiang, China, [2]Faculty of Education, Henan University, Kaifeng, China, [3]Hongqi District Second Experimental Primary School, Xinxiang, China

Interpersonal trust is the premise and foundation of encouraging cooperation in this age of rapid progress. The purpose of this study was to investigate how moral judgment affects bystanders' interpersonal trust and its internal mechanisms when there are ethical transgressions. The moral judgment of the evaluators was divided into three categories—opposition, neutrality and approval—on the basis of the moral transgressions of the offenders. Three moral judgment circumstances were randomly assigned to 143 primary school pupils, and the assessors scored the children via trustworthiness and trust scales. According to the findings, interpersonal trust is significantly predicted by moral judgment. Compared with neutral judgment, opposing moral violations significantly improves bystanders' interpersonal trust in the evaluator, whereas approving moral violations does not significantly predict interpersonal trust. Trustworthiness plays a mediating role in the influence of moral judgment on interpersonal trust. Compared with neutral judgment, trustworthiness mediates the influence of opposed judgment on interpersonal trust rather than the influence of approved judgment on interpersonal trust. The findings demonstrate that moral opposition to transgressions influences interpersonal trust either directly or indirectly through trustworthiness.

KEYWORDS

interpersonal trust, moral judgment, trustworthiness, mediation effect, adolescent

1 Introduction

Interpersonal trust is the cornerstone of human cooperation and collaboration, permeating all aspects of social life. As a lubricant for social interaction, trust not only plays a crucial role in initiating, establishing, and maintaining intimate relationships, but also contributes to the prosperity of groups, organizations, and nations (Dunning et al., 2019; Weiss et al., 2021). Over the past two decades, the evolution and impact factors of trust behaviors in humans and other populations have become a central topic of discussion in fields such as economics (Chetty et al., 2021), politics (Carlin et al., 2022), psychology (Weiss et al., 2021), and cognitive neuroscience (Krueger and Meyer-Lindenberg, 2019), and have been extensively explored. By means of economic game tasks and self-report questionnaires, a large number of studies have discovered that trusting others is a ubiquitous social preference (Dunning et al., 2014), possesses a certain degree of biological heritability (Riedl and Javor, 2012), and is prone to the interactive influence of personality and situational factors (Weinschenk and Dawes, 2019).

In the context of interpersonal interactions, individuals frequently depend on a range of social cues—such as facial characteristics, reputation information, and observable behaviors—to assess the trustworthiness of others and ultimately determine whether to

extend their trust. Costly signaling theory posits that expensive social behaviors, like third-party interventions, indicate trustworthiness to bystanders, influencing their interpersonal trust toward to the interveners (BliegeBird and Smith, 2005; Gintis et al., 2001). Third-party punishers are viewed as more trustworthy in economic game tasks than persons who do not carry out punishment, according to Jordan et al. (2016). Other studies have found that corrupt third parties undermine trust and prosocial behavior between people (Spadaro et al., 2023). Sun et al. (2023) also found that third-party punishment affected bystanders'trust in punishers in both the in-group and out-group conditions. In addition, studies have analyzed the process by which punishment systems shape trust (Olcina and Calabuig, 2021). However, there are not many work exploring the impact of non-third party interventions on interpersonal trust. Moral judgment is a way of non-third-party intervention, and will it have a similar impact on interpersonal trust?

Moral judgment is a crucial way for individuals to intervene in moral transgressions (Lergetporer et al., 2014). As a type of social cue, moral judgment can convey information such as the trustworthiness of the person who is assigning judgment (Connelly et al., 2011; Haidt, 2001). Nonetheless, there are certain limitations in earlier studies on moral judgment and interpersonal trust. First, previous studies have discussed separately two types of moral judgment, disapproval and approval, but few studies have examined and compared the two together (Bostyn et al., 2023; Everett et al., 2016); however, simultaneously exploring the relationship between the two and interpersonal trust plays an important role in motivating and strengthening human cooperation (Guglielmo and Malle, 2019). Second, the behavioral game task—which is less common in everyday life—was applied in the majority of earlier studies (Karlan, 2005; Thielmann and Hilbig, 2015), thus limiting the generalizability of the findings to real-world contexts. Third, while earlier research has been conducted on adult populations, it is crucial to concentrate on adolescents to understand how the relationship between moral judgment and interpersonal trust develops because adults and adolescents have different age features (Bostyn and Roets, 2017; Everett et al., 2021; Towner et al., 2023). Adolescence is an important period for the development of individual moral cognition. According to the Kohlberg's stages of moral development, students in this period were in the stage of seeking directional recognition, and their moral values were oriented by interpersonal harmony (Walker, 1982). The study of the relationship between the moral judgment and the interpersonal trust during this period can help them to establish a good peer relationship and promote the cooperation and development of the future society. Moreover, previous studies mostly focus on adult groups, and the socialization degree of adult groups is much higher than that of adolescent (Simpson et al., 2013; Bostyn and Roets, 2017; Bostyn et al., 2023; Everett et al., 2016). The study of adolescent is helpful to clarify the development process of moral judgment affecting the interpersonal trust of bystanders and enrich the content of this field.

1.1 Moral judgment and interpersonal trust

Human society is constrained by various moral norms. Social moral norms can be effectively upheld, and social justice and fairness can be promoted by individual intervention in moral transgressions. The term "moral judgment" primarily refers to the perceiver's assessment of a breach of moral standards, and it may be classified into four types: evaluation judgment, normative judgment, moral error judgment, and blame judgment (Malle, 2021). According to costly signaling theory, opposing moral violations can send a signal to others that people may trust the evaluator more in cases of risk and uncertainty (BliegeBird and Smith, 2005). Approving moral transgressions is contrary to modern society's norms, but it can also send a message to others that the evaluator is not someone to trust on a personal level (Gintis et al., 2001). Moreover, it has been demonstrated that a person's moral judgment of immoral activity increases his or her degree of trust (Simpson et al., 2013). Kennedy and Schweitzer (2018) also reported that when accusing others of immoral behavior, the individual sends a signal of his or her moral character, thus increasing the other person's interpersonal trust. Other studies have shown that a person's approved judgment of immoral activity can negatively affect the perception of bystanders and reduce their level of trust is that individual (Uhlmann et al., 2013; Bostyn and Roets, 2017). Therefore, Hypothesis 1 is as follows: Interpersonal trust is impacted by moral judgment. Compared with neutral judgment, opposing moral violations positively predicts interpersonal trust, and approving moral violations negatively predicts interpersonal trust.

1.2 Trustworthiness as a potential mediator

Trustworthiness is the perception of qualities such as individual ability, benevolence, and integrity (Colquitt et al., 2007; Mayer et al., 1995). On the one hand, studies have indicated that observers deduce an individual's personality from the moral judgments made by others (Kreps and Monin, 2014; Sacco et al., 2017; Uhlmann et al., 2013). The costly signaling theory also holds that an individual's judgment of immoral behavior sends a signal that indicates the individual's ability, benevolence and integrity, which constitute the perception of the individual's trustworthiness (Gintis et al., 2001; Mayer et al., 1995). In other words, moral judgment affects trustworthiness. Research has demonstrated that moral judgment has an impact on trustworthiness and that people tend to view strong moral judges as more trustworthy than weak moral judges (Simpson et al., 2013). According to Everett et al. (2016), moral judgment can be used to infer an individual's trustworthiness. Therefore, moral judgment can play a role in establishing trustworthiness.

On the other hand, according to the ABI model proposed by Mayer et al. (1995), trust can be examined from three perspectives—ability, benevolence and integrity—in which ability and integrity contribute to cognition-based trust (McAllister, 1995) whereas benevolence contributes to emotion-based trust (Dirks and Ferrin, 2002). Some studies have also shown that characteristics such as perceived individual ability, benevolence and integrity have a direct effect on the degree to which others trust a person (Sun et al., 2023; Wang and Murnighan, 2017). All of these studies highlight the importance of trustworthiness in building interpersonal trust. Kennedy and Schweitzer (2018) also reported that the perception of integrity plays a mediating role between alleging unethical behavior and interpersonal trust. Therefore, Hypothesis 2 is as follows: Trustworthiness plays a

mediating role in the influence of moral judgment on interpersonal trust. Compared with neutral people, opposing moral violations increases interpersonal trust by increasing the perception of trustworthiness, and approving moral violations reduces interpersonal trust by reducing that perception.

To test the above hypothesis, this study improved upon previous studies by using adolescents as the subjects and proposed the use of a school class moral violation scenario to explore the influence of moral judgment on interpersonal trust and the mediating role of trustworthiness. This study largely uses a paper experiment with a single-factor, three-level interexperimental design to investigate the relationships among moral judgment, trustworthiness, and interpersonal trust. The three moral judgments are opposed, neutral, and approved.

2 Method

2.1 Participants

A total of 162 questionnaires were collected from a primary school in Henan Province, among which 19 questionnaires were excluded because of missing items. Ultimately, 143 questionnaires were valid, and the effective response rate of the questionnaires was 88.27%. Our sample size was determined through an *a priori* power analysis, assuming an α of 0.05 and a power of 0.80, indicating that the minimum effect size we had power to detect was a medium effect of $f = 0.25$ (Faul et al., 2007). The subjects' ages ranged from 11 to 13 years, with an average age of 11.29 years ($SD = 0.50$). Of the 143 subjects, 69 (48.30%) were female, and 74 (51.70%) were male. All the subjects were physically and mentally healthy, had no history of mental illness, were right-handed, and had normal or corrected-to-normal vision. Ethics committee approval was obtained from the Faculty of Education at Henan Normal University, and protocol adherence to the Declaration of Helsinki was ensured.

2.2 Experimental materials

2.2.1 Moral violation scenario

The moral violation scenario employed in this study was adapted from Dhaliwal et al. (2020). "Imagine this scene: classmate A was loud in front of all classmates and teachers during a class meeting last week." This was the exact description of the class violation used in this study. The three types of moral judgments were as follows: approved (Monitor B praises classmate A and believes it is morally appropriate), neutral (Monitor B does not condemn or praise classmate A), and opposed (Monitor B condemns classmate A and considers it morally inappropriate). Four students were asked to participate in a pretest to ensure that the subjects grasped the material and the situation before the official test. After the experiment, the subjects were interviewed, and all the subjects correctly understood the situation and related questions.

2.2.2 Trustworthiness scale

This scale is adapted from the trustworthiness scale proposed by Mayer and Davis (1999). The scale is divided into three dimensions—ability, integrity and benevolence—with a total of 17 items. We changed "top management" in the original scale to "monitor B" and "work" to "class work." All other parts remained unchanged. Questions 1–6 measure ability (e.g., "Monitor B is very capable of performing class work"); questions 7–11 measure benevolence (e.g., "My needs and wishes are very important to monitor B"); and questions 12–17 measure integrity (e.g., "Monitor B has a strong sense of justice"). The scale was rated on a 5-point Likert scale, where 1 represented complete disagreement and 5 represented complete agreement. Higher scores on the scale denoted greater trustworthiness. The internal consistency coefficient of this scale in this study was 0.92, and the internal consistency coefficients of the ability, benevolence and integrity dimensions were 0.90, 0.81, and 0.73, respectively. We performed a confirmatory factor analysis of this scale, and the results are as follows: $\chi^2 = 190.73$, $df = 114$, $\chi^2/df = 1.67$ ($p < 0.001$), TLI = 0.93, CFI = 0.94, RMSEA = 0.07, and SRMR = 0.05.

2.2.3 Trust scale

The trust scale used in the Ng and Chua study contains 8 items, in which questions 1--4 measure cognitive trust and questions 5–8 measure emotional trust (Ng and Chua, 2006). We have adapted this scale. We changed "they" in the original scale to "monitor B" and changed "teamwork" to "class work." A representative item for measuring cognitive trust is "Monitor B is the person who takes class work seriously." A representative item for measuring emotional trust is "You can freely talk to monitor B about your difficulties in learning and know that monitor B is willing to listen." The scale was rated on a 5-point Likert scale, where 1 represented complete disagreement and 5 represented complete agreement. The internal consistency coefficient of this scale in this study was 0.88, and the internal consistency coefficients of the cognitive trust and emotional trust subscales were 0.91 and 0.80, respectively. We performed a confirmatory factor analysis of this scale, and the results are as follows: $\chi^2 = 19.62$, $df = 11$, $\chi^2/df = 1.78$ ($p < 0.001$), TLI = 0.97, CFI = 0.99, RMSEA = 0.07, and SRMR = 0.04.

3 Results

3.1 Descriptive statistics and correlation analysis

Table 1 displays the descriptive statistics and correlation analysis results for moral judgment, trustworthiness and interpersonal trust. The results of the correlation analysis reveal that moral judgment was significantly negatively correlated with both trustworthiness and interpersonal trust and that trustworthiness was significantly positively correlated with interpersonal trust.

TABLE 1 Descriptive statistics and correlation analysis results ($N = 143$).

	M (*SD*)	1	2	3
1 Moral judgment	2.03 (0.80)	1		
2 Trustworthiness	2.91 (0.94)	−0.54***	1	
3 Interpersonal trust	2.91 (1.08)	−0.55***	0.87***	1

Moral judgment: Opposed judgment = 1, Neutral judgment = 2, Approved judgment = 3.
***$p < 0.001$.

3.2 Preliminary analyses

One-way ANOVA was conducted with moral judgment as the independent variable and trustworthiness and interpersonal trust as the dependent variables. The results are shown in Figure 1. When trustworthiness was used as the dependent variable, the main effect of moral judgment was significant, $F = 32.99$, $p < 0.001$. Multiple comparisons reveal that the effect in the opposed group ($M = 3.70$, $SD = 0.65$) was significantly greater than that in the neutral group ($M = 2.71$, $SD = 0.70$) and the approved group ($M = 2.43$, $SD = 0.95$), whereas there was no significant difference between the neutral and approved groups. When interpersonal trust was used as the dependent variable, the main effect of moral judgment was significant ($F = 44.99$, $p < 0.001$). Multiple comparisons reveal that the effect in the opposed group ($M = 3.94$, $SD = 0.62$) was significantly greater than that in the neutral group ($M = 2.53$, $SD = 0.87$) and the approved group ($M = 2.41$, $SD = 1.00$), whereas there was no significant difference between the neutral and approved groups.

3.3 Mediation model

The mediating effect of trustworthiness between moral judgment and interpersonal trust was examined via SPSS 25.0 and Mplus 8.3. Moral judgment is dummy coded prior to the mediation effect analysis because it is a three-categorical variable. The neutral judgment group was set as the reference group to further reveal the causal relationship between moral judgment and interpersonal trust by comparison with the opposed judgment group and the approved judgment group. The dependent variable was interpersonal trust. All variables were normalized prior to the examination of the mediating effect. The results of the regression analysis are presented in Table 2.

According to the regression coefficient of the regression equation with moral judgment as the predictor variable, interpersonal trust as the outcome variable and the results of the significance test, opposed judgment (vs. neutral judgment) can significantly positively predict interpersonal trust ($\beta = 0.60$, $p < 0.001$), but approved judgment (vs. neutral judgment) cannot significantly predict interpersonal trust. According to the regression coefficient of the regression equation with moral judgment as the predictor variable, trustworthiness as the outcome variable and the results of the significance test, opposed judgment (vs. neutral judgment) can significantly positively predict trustworthiness ($\beta = 0.48$, $p < 0.001$), but approved judgment (vs. neutral judgment) cannot significantly predict trustworthiness. The regression coefficient and significance test results show that when moral judgment and trustworthiness both affect interpersonal trust, opposed judgment (vs. neutral judgment) can significantly positively predict interpersonal trust ($\beta = 0.23$, $p < 0.001$), approved judgment (vs. neutral judgment) cannot significantly predict interpersonal trust, and trustworthiness can significantly positively predict interpersonal trust ($\beta = 0.76$, $p < 0.001$).

This study tested the mediating effect of trustworthiness via a structural equation model to further identify the mediating role. The fit indices of the structural equation model are represented by χ^2, CFI, TLI, RMSEA and SRMR. On the basis of this analysis, the results indicate that while the model fit is acceptable given the study's context, it is not ideal, with $\chi^2 = 54.22$, $df = 18$, $\chi^2/df = 3.01$ ($p < 0.001$), TLI = 0.88, CFI = 0.94, RMSEA = 0.12, and SRMR = 0.06. The two latent variables are trustworthiness, i.e., ability, benevolence and integrity; and interpersonal trust, i.e., cognitive trust and emotional trust, with the neutral judgment group serving as the reference variable and the independent variables being coded as virtual variables. With trustworthiness serving as the mediating variable and interpersonal trust serving as the dependent variable, the results are shown in Figure 2. The analysis with trustworthiness serving as the mediating variable reveals that the total effect of the opposed judgment group was significant ($c_1 = 0.77$, $p < 0.001$), the direct effect was significant ($c'_1 = 0.20$, $p < 0.05$), and the indirect effect through trustworthiness was also significant ($ab_1 = 0.57$, 95% CI [0.41, 0.77]). However, the total effect, direct effect, and indirect effect through trustworthiness of the approved judgment group were not significant.

4 Discussion

This study indicates that moral judgment can influence bystanders' perceptions of trustworthiness and interpersonal trust. This is in line with previous studies (Simpson et al., 2013). For trustworthiness and interpersonal trust, this study revealed that different moral judgments can lead to different levels of trustworthiness perceptions and interpersonal trust. Specifically, the perception of trustworthiness and interpersonal trust of evaluators who make opposed judgments is greater than that of evaluators who make neutral judgments and approved judgments, whereas there is no significant difference in the perceived trustworthiness and interpersonal trust levels between evaluators who make neutral judgments and those who make approved judgments.

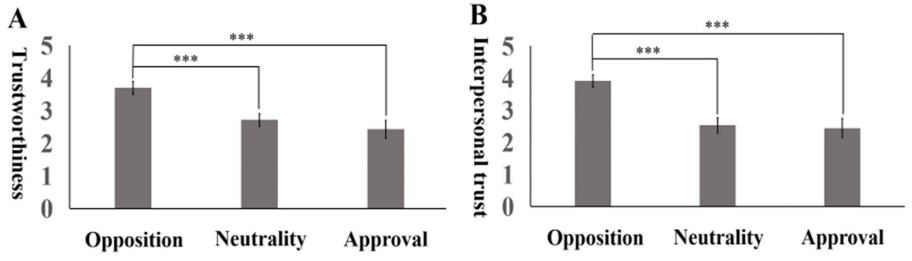


FIGURE 1
Differences in the effects of moral judgment on trustworthiness **(A)** and interpersonal trust **(B)**. ***$p < 0.001$.

TABLE 2 Regression analysis of moral judgment, trustworthiness and interpersonal trust.

Regression equation		Overall fit index			Significance of the regression coefficient	
Outcome variable	Predictor variable	R	R^2	F	β	t
Interpersonal trust	Opposed judgment (vs. neutral judgment)	0.63	0.39	44.99***	0.60	8.06***
	Approved judgment (vs. neutral judgment)				−0.05	−0.67
Trustworthiness	Opposed judgment (vs. neutral judgment)	0.57	0.32	32.99***	0.48	6.14***
	Approved judgment (vs. neutral judgment)				−0.15	−1.84
Interpersonal trust	Opposed judgment (vs. neutral judgment)	0.89	0.79	169.57***	0.23	4.64***
	Approved judgment (vs. neutral judgment)				0.06	1.34
	Trustworthiness				0.76	15.98***

All variables in the model are brought back into the equation after standardized processing. ***$p < 0.001$.

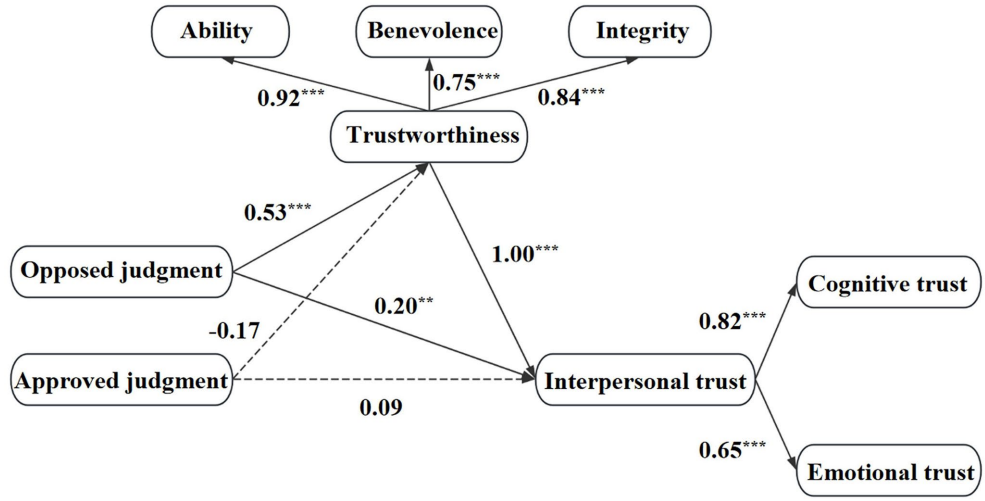


FIGURE 2
Mediating effect diagram of trustworthiness. ***$p < 0.001$.

Opposition to violators in the face of moral transgressions highlight the values and positions of the evaluator. People are more likely to positively identify with an evaluator and view them as trustworthy when they believe that the evaluator shares their same values or the recognized values of their social group (Connelly et al., 2011; Haidt, 2001). This phenomenon also increases people's interpersonal trust. Moreover, costly signaling theory suggests that opposing moral violations sends a signal to bystanders that the evaluator is trustworthy, thereby suggesting that the evaluator is more likely to be trusted in cases of risk and uncertainty (BliegeBird and Smith, 2005; Gintis et al., 2001). Indirect reciprocity theory also states that opposed judgments emphasize that moral violations are incorrect, prevent violators from harming others, safely safeguard the interests of others, maintain fairness and order within the group, and conclude that people are more likely to form positive cognitions of such assessors (Nowak and Sigmund, 2005). Therefore, people will trust evaluators who make opposed judgments more than those who make neutral and approved judgments. Moreover, it is possible that the study's relatively minor moral violations have less severe consequences, so there is no significant difference in the perceived trustworthiness and interpersonal trust levels between the evaluators who make neutral judgments and those who make approved judgments.

In this study, trustworthiness plays a partial mediating role in the effect of moral judgment on interpersonal trust. That is, moral judgment affects the interpersonal trust of bystanders by influencing their perception of trustworthiness. A finding that is consistent with previous findings (Kennedy and Schweitzer, 2018; Jordan et al., 2016; Simpson et al., 2013). Opposed judgments can prevent such behavior in the future, safely safeguard the interests of others, maintain fairness and order within the group, and highlight the individual's own moral standards by indicating the individual's attitude toward moral violations. Opposed judgment also reveals the individual's qualities of ability, benevolence, and integrity, and the perception of these individual qualities is actually the perception of the individual's trustworthiness (Mayer et al., 1995). Simply stated, individuals who oppose moral violations are considered trustworthy. Trustworthiness is the premise and basis of interpersonal trust (Sun et al., 2023; Wang and Murnighan, 2017), which is composed of both cognitive and emotional trust. The perception of ability and integrity contributes to cognitive-based trust (McAllister, 1995), and the perception of benevolence contributes to emotion-based trust (Dirks and Ferrin, 2002). Therefore, in contrast to neutral judgments, opposed judgments not only directly affect interpersonal trust but also indirectly affect interpersonal trust by affecting trustworthiness. That is,

trustworthiness plays a partial mediating role. In the moral violation scenario involved in this study, the intention to oppose moral violation is clear, and the intention to approve moral violation is vague (Carlson et al., 2022). Individuals may actually approve of such moral violations, or they may contrarily approve them for other reasons, such as pressure from peer relationships. Moreover, owing to the low degree of socialization of adolescents during this period, it may not be easy to associate approved judgments with individual moral qualities, so approved judgments do not predict well bystanders' interpersonal trust. In addition, because the moral violation scenario used in this study was in a school class and the consequences of the violation were relatively light, there was no significant difference in the perceptions of trustworthiness and interpersonal trust of the assessors who made the neutral and approved judgments. Therefore, compared with neutral judgment, trustworthiness plays no mediating role in the influence of approved judgment on interpersonal trust. The costly signaling theory also supports this result, whereas an opposed judgment requires individuals to pay the corresponding cost and bear the risk of retaliation by the violators, which sends a signal to the bystander that the evaluator is trustworthy and increases bystanders' interpersonal trust. An approved judgment may send a weak signal to bystanders, causing trustworthiness to play no role in the impact of an approved judgment on bystanders' interpersonal trust (BliegeBird and Smith, 2005; Gintis et al., 2001).

4.1 Implications of the study

This study has important theoretical significance and practical value. With respect to theoretical significance, previous studies have focused more on the influence of moral judgment on the evaluator himself or herself. This study expands on this to verify the spillover effect of moral judgment on the interpersonal trust of bystanders. Moreover, this study, which was conducted in the context of an Eastern culture, differs culturally from most studies conducted in Western cultures, and culture is an important factor affecting individual cognition and behavior. Accordingly, this study once again demonstrates the relationship between moral judgment and bystanders' interpersonal trust from a cross-cultural perspective. In terms of practical value, this study revealed that individual moral judgments directly affect the interpersonal trust of bystanders, which may motivate individuals to make prudent moral judgments in public to win the trust of more bystanders. Furthermore, this study revealed that moral judgment indirectly influences interpersonal trust by influencing bystanders' perceptions of trustworthiness, which then strengthens the importance of trust building qualities such as ability, benevolence and integrity. These findings contribute to the individual's understanding of the process of building trust and show how to demonstrate their good qualities through moral judgment in social interactions to improve their reliability and gain the trust of others.

4.2 Limitations and future research

Although this study reveals the development mechanism of interpersonal trust during the process of social interaction and the mediating role of trustworthiness, there are several limitations. First, the sample size of this study was not very representative. The sample size was selected based on some criteria, which restricts the sample to a more specific population. Future studies could consider reducing the number of inclusion criteria and generalizing the findings to a broader population. Second, the monitor, who has more rights and responsibilities than do ordinary students, made moral judgments in this study, thus implying that the monitor's moral judgment had greater authority. In the future, whether there exists such a relationship between the moral judgment of ordinary students and interpersonal trust should be further examined. Third, given that people's attitudes and behaviors are not always consistent, future research should consider using behavioral paradigms instead of the scales used in this study to measure the behavioral performance of interpersonal trust. Finally, the relationship between moral judgment and interpersonal trust under varying levels of moral violation can be investigated in the future to improve the external validity of such studies, as the degree of moral violation in the scenario used in this study was relatively small, whereas the degrees of moral violations in real life are not equal.

Moreover, this study has important implications for educational practice. In the face of moral violations among students, individuals with influential roles, such as teachers and monitors, should help students make correct moral judgments in a timely manner, which, in turn, improves their (the teachers and monitors) level of trustworthiness and helps them to better manage student behavior. In addition, when providing moral education programs for students, schools should focus on the value of moral knowledge, encourage students to establish correct moral values, and support students as they render moral judgments on moral violations. Taken together, these actions will improve interpersonal trust among students and promote win–win cooperation among young people.

5 Conclusion

This study explored the influence of moral judgment on interpersonal trust and its underlying mechanisms via structural equation modeling. Two significant conclusions were reached.(1) Bystanders' interpersonal trust in the moral evaluator is impacted by the bystanders' moral judgment of moral transgressions. Compared with people who make neutral and approved judgments, bystanders have greater trust in people who make opposed judgment. (2) Trustworthiness plays a mediating role in the influence of moral judgment on interpersonal trust; that is, moral judgment affects bystanders' perceptions of trustworthiness and subsequently affects bystanders' interpersonal trust.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Faculty of Education at Henan Normal University. The studies were conducted

in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

ZhZ: Conceptualization, Funding acquisition, Writing – review & editing. XC: Data curation, Formal analysis, Writing – original draft. WG: Data curation, Formal analysis, Writing – original draft. ZeZ: Data curation, Formal analysis, Writing – original draft. CQ: Conceptualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Social Science Foundation of China [24BSH105].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

BliegeBird, R., and Smith, E. (2005). Signaling theory, strategic interaction, and symbolic capital. *Curr. Anthropol.* 46, 221–248. doi: 10.1086/427115

Bostyn, D. H., Chandrashekar, S. P., and Roets, A. (2023). Deontologists are not always trusted over utilitarians: revisiting inferences of trustworthiness from moral judgments. *Sci. Rep.* 13:1665. doi: 10.1038/s41598-023-27943-3

Bostyn, D. H., and Roets, A. (2017). Trust, trolleys and social dilemmas: a replication study. *J. Exp. Psychol. Gen.* 146, e1–e7. doi: 10.1037/xge0000295

Carlin, R. E., González, R., Love, G. J., Miranda, D. A., and Navia, P. D. (2022). Ethnicity or policy? The conditioning of intergroup trust in the context of ethnic conflict. *Polit. Psychol.* 43, 201–220. doi: 10.1111/POPS.12747

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., and Crockett, M. J. (2022). How inferred motives shape moral judgements. *Nat. Rev. Psychol.* 1, 468–478. doi: 10.1038/s44159-022-00071-x

Chetty, R., Hofmeyr, A., Kincaid, H., and Monroe, B. (2021). The trust game does not (only) measure trust: the risk-trust confound revisited. *J. Behav. Exp. Econ.* 90:101520. doi: 10.1016/j.socec.2020.101520

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* 92, 909–927. doi: 10.1037/0021-9010.92.4.909

Connelly, B. L., Certo, S. T., Ireland, R. D., and Reutzel, C. R. (2011). Signaling theory: a review and assessment. *J. Manag.* 37, 39–67. doi: 10.1177/0149206310388419

Dhaliwal, N. A., Skarlicki, D. P., Hoegg, J., and Daniels, M. A. (2020). Consequentialist motives for punishment signal trustworthiness. *J. Bus. Ethics* 176, 451–466. doi: 10.1007/s10551-020-04664-5

Dirks, K. T., and Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *J. Appl. Psychol.* 87, 611–628. doi: 10.1037/0021-9010.87.4.611

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., and Fetchenhauer, D. (2014). Trust at zero acquaintance: more a matter of respect than expectation of reward. *J. Pers. Soc. Psychol.* 107, 122–141. doi: 10.1037/a0036673

Dunning, D., Fetchenhauer, D., and Schlösser, T. (2019). Why people trust: solved puzzles and open mysteries. *Curr. Dir. Psychol. Sci.* 28, 366–371. doi: 10.1177/0963721419838255

Everett, J. A., Colombatto, C., Awad, E., Boggio, P., Bos, B., Brady, W. J., et al. (2021). Moral dilemmas and trust in leaders during a global health crisis. *Nat. Hum. Behav.* 5, 1074–1088. doi: 10.1038/s41562-021-01156-y

Everett, J. A. C., Pizarro, D. A., and Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* 145, 772–787. doi: 10.1037/xge0000165

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Gintis, H., Smith, E. A., and Bowles, S. (2001). Costly signaling and cooperation. *J. Theor. Biol.* 213, 103–119. doi: 10.1006/jtbi.2001.2406

Guglielmo, S., and Malle, B. F. (2019). Asymmetric morality: blame is more differentiated and more extreme than praise. *PLoS One* 14:e0213544. doi: 10.1371/journal.pone.0213544

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814–834. doi: 10.1037/0033-295X.108.4.814

Jordan, J. J., Hoffman, M., Bloom, P., and Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature* 530, 473–476. doi: 10.1038/nature16981

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *Am. Econ. Rev.* 95, 1688–1699. doi: 10.1257/000282805775014407

Kennedy, J. A., and Schweitzer, M. E. (2018). Building trust by tearing others down: when accusing others of unethical behavior engenders trust. *Organ. Behav. Hum. Decis. Process.* 149, 111–128. doi: 10.1016/j.obhdp.2018.10.001

Kreps, T. A., and Monin, B. (2014). Core values versus common sense: consequentialist views appear less rooted in morality. *Personal. Soc. Psychol. Bull.* 40, 1529–1542. doi: 10.1177/0146167214551154

Krueger, F., and Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends Neurosci.* 42, 92–101. doi: 10.1016/j.tins.2018.10.004

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., and Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proc. Natl. Acad. Sci. USA* 111, 6916–6921. doi: 10.1073/pnas.1320451111

Malle, B. F. (2021). Moral judgments. *Annu. Rev. Psychol.* 72, 293–318. doi: 10.1146/annurev-psych-072220-104358

Mayer, R. C., and Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: a field quasi-experiment. *J. Appl. Psychol.* 84, 123–136. doi: 10.1037/0021-9010.84.1.123

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.2307/258792

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Acad. Manag. J.* 38, 24–59. doi: 10.2307/256727

Ng, K. Y., and Chua, R. Y. (2006). Do I contribute more when I trust more? Differential effects of cognition-and affect-based trust. *Manag. Organ. Rev.* 2, 43–66. doi: 10.1111/j.1740-8784.2006.00028.x

Nowak, M. A., and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature* 437, 1291–1298. doi: 10.1038/nature04131

Olcina, G., and Calabuig, V. (2021). Trust and punishment. *Eur. J. Polit. Econ.* 70:102032. doi: 10.1016/j.ejpoleco.2021.102032

Riedl, R., and Javor, A. (2012). The biology of trust: integrating evidence from genetics, endocrinology, and functional brain imaging. *J. Neurosci. Psychol. E* 5, 63–91. doi: 10.1037/a0026318

Sacco, D. F., Brown, M., Lustgraaf, C. J., and Hugenberg, K. (2017). The adaptive utility of deontology: deontological moral decision-making fosters perceptions of trust and likeability. *Evol. Psychol. Sci.* 3, 125–132. doi: 10.1007/s40806-016-0080-6

Simpson, B., Harrell, A., and Willer, R. (2013). Hidden paths from morality to cooperation: moral judgments promote trust and trustworthiness. *Soc. Forces* 91, 1529–1548. doi: 10.1093/sf/sot015

Spadaro, G., Molho, C., Van Prooijen, J. W., Romano, A., Mosso, C. O., and Van Lange, P. A. (2023). Corrupt third parties undermine trust and prosocial behaviour between people. *Nat. Hum. Behav.* 7, 46–54. doi: 10.1038/s41562-022-01457-w

Sun, B., Jin, L., Yue, G., and Ren, Z. (2023). Is a punisher always trustworthy? In-group punishment reduces trust. *Curr. Psychol.* 42, 22965–22975. doi: 10.1007/s12144-022-03395-2

Thielmann, I., and Hilbig, B. E. (2015). Trust: an integrative review from a person–situation perspective. *Rev. Gen. Psychol.* 19, 249–277. doi: 10.1037/gpr0000046

Towner, E., Chierchia, G., and Blakemore, S. J. (2023). Sensitivity and specificity in affective and social learning in adolescence. *Trends Cogn. Sci.* 27, 642–655. doi: 10.1016/j.tics.2023.04.002

Uhlmann, E. L., Zhu, L. L., and Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition* 126, 326–334. doi: 10.1016/j.cognition.2012.10.005

Walker, L. J. (1982). The sequentiality of Kohlberg's stages of moral development. *Child Dev.* 53, 1330–1336. doi: 10.1111/j.1467-8624.1982.tb04172.x

Wang, L., and Murnighan, J. K. (2017). The dynamics of punishment and trust. *J. Appl. Psychol.* 102, 1385–1402. doi: 10.1037/apl0000178

Weinschenk, A. C., and Dawes, C. T. (2019). The genetic and psychological underpinnings of generalized social trust. *J. Trust Res.* 9, 47–65. doi: 10.1080/21515581.2018.1497516

Weiss, A., Michels, C., Burgmer, P., Mussweiler, T., Ockenfels, A., and Hofmann, W. (2021). Trust in everyday life. *J. Pers. Soc. Psychol.* 121, 95–114. doi: 10.1037/pspi0000334

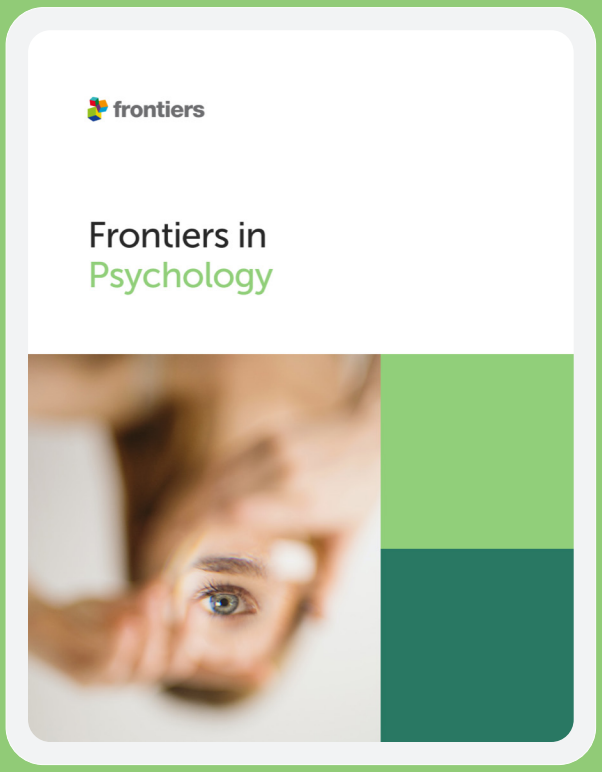
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact