ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202408.00001

Structure and Function of Harm Aversion Based on the Dissociation Paradigm

Authors: Cen Yushan, Xia Lingxiang, Huang Runyu, Lu Jie, Xia Lingxiang

Date: 2024-07-24T00:00:00+00:00

Abstract

Harm aversion is central to human morality, yet the structure of harm aversion and its mechanisms for inhibiting aggression remain unclear. Therefore, one pilot experiment and three formal experiments were conducted. The pilot experiment developed a harm behavior/outcome aversion dissociation task based on the process dissociation procedure. Experiment 1 employed this task to examine the structure of harm aversion, while Experiment 2 (including 2a and 2b) investigated the moral cognitive mechanisms through which harm behavior/outcome aversion inhibits aggression. Results demonstrated that harm aversion comprises two independent factors: harm behavior aversion and harm outcome aversion. Harm outcome aversion can inhibit aggression through moral disengagement, whereas the inhibitory effects of harm behavior aversion on aggression and moral disengagement were not robust. These experiments developed research tools for harm aversion, examined its two-factor structure and differential effects on inhibiting aggressive cognition and behavior, revealed the moral cognitive pathways through which harm aversion inhibits aggression, and advanced the theory of moral emotions and aggression.

Full Text

Preamble

Self-Check Report for *Acta Psychologica Sinica*
Please complete the following items and paste them on the first page of your manuscript.

1. Please list up to three innovative contributions of this study in the form of "Research Highlights" (must include theoretical contributions), with a total of no more than 200 words.

The goal of *Acta Psychologica Sinica* is to publish cutting-edge psychological research that is "both scientifically excellent and of particularly broad interest and significance." If your study only makes minor incremental contributions, does not attempt to open new areas of inquiry, or fails to propose unique and innovative perspectives—especially if it is purely an algorithm or technical work without clear psychological research questions—such studies have a low chance of acceptance. We recommend submitting to other journals.

Answer: First, this study reveals the relative independence of two components of harm aversion (harm action aversion and harm outcome aversion) and their differential effects in inhibiting aggressive cognition and behavior. Second, it demonstrates that moral disengagement serves as an important moral cognitive pathway through which harm aversion inhibits aggression. Third, based on the process dissociation procedure, we developed a research tool for separating harm action/outcome aversion.

2. Have you used the same data in other submitted or published articles? If yes, please attach those articles for review. (We do not encourage authors to publish multiple articles using the same data with identical variables, nor do we support splitting a series of related studies into multiple publications.)

Answer: The data used in this study are independent from those used in other papers, with no instances of data reuse.

3. Non-experimental, non-intervention studies in management, clinical, personality, and social psychology that rely solely on self-report (questionnaire) methods need to examine common method bias. What methods did you use to control for or demonstrate that such bias does not affect the validity of your conclusions? (For literature on common method bias, see: http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract894.shtml) Studies based on cross-sectional data with only self-reports and convenient samples are easy to conduct but typically lack innovative value and have low acceptance rates.

Answer: All four studies in this paper are experimental studies based on a dissociation paradigm, so common method bias is not an issue.

4. Did you report and analyze effect sizes (e.g., Cohen's d for t-tests, ^{2} or ^{2}p for ANOVA, standardized regression coefficients)? (Many studies mechanically report effect sizes without necessary analysis or explanation, such as whether the effect size is small, medium, or large, or its theoretical/applied significance.) (Search "effect size calculator" on Google for convenient apps. For explanations of effect sizes in Chinese, see: http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract1150.shtml; in English, see: http://www.uccs.edu/lbecker/effect-size.html) Did you report 95% CIs for statistical analyses? (e.g., 95% CI for differences, correlations/regression coefficients) For calculations and graphing of confidence intervals, see https://thenewstatistics.com/itns/esci/

Answer: In all three formal experiments, we used G*Power to calculate sample size, setting statistical power at 85%. Specifically, Experiment 1 had a planned sample size of 218, with an effective sample of 287; Experiment 2a had a planned sample size of 424 (due to questionnaire measurement of aggression and moral disengagement, we set a small-to-medium effect size $f^2 = 0.07$), with an effective sample of 423; Experiment 2b had a planned sample size of 102 (due to aggression task measurement of moral disengagement and aggressive behavior, we set a medium effect size $f^2 = 0.15$), with an effective sample of 107. The effective samples were generally consistent with the planned sample sizes.

6. For hypothesis testing using null hypothesis significance testing (NHST), precise p-values should be reported rather than p-value ranges (report ranges only for $p < 0.001$, otherwise report exact p-values). Does your paper meet this requirement? For Bayesian factors, have you reported their sensitivity to prior distribution assumptions?

Answer: In this paper, all analyses report exact p-values.

7. To ensure completeness of data reporting, if any data were excluded from statistical analyses, were they reported in the text? What were the reasons? How would the results change if those data were included? How were missing data handled in statistical analyses? When using scales, were any individual items deleted? Why? How would the results change if those items were included? Were there any measured items or variables not reported? Why? Please indicate where in the paper this is addressed.

Answer: First, excluded participants were reported in the text. Specifically, exclusions were based on two main criteria: (1) participants who failed validity checks or completed questionnaires/experiments too quickly were removed; (2) according to the process dissociation procedure (PDP) rules (Conway & Gawronski, 2013; Du & Liu, 2018), participants who consistently chose not to harm in congruent conditions were deleted. Second, this paper used maximum likelihood estimation for missing data. Third, complete scales were used without deleting any items. Fourth, in Experiment 2a, we also measured reactive-proactive aggression but did not report it. This variable was initially included to retest the validity of our developed "harm action/outcome aversion dissociation task." Although results supported our hypothesis, reviewers noted during preliminary review that this test was unnecessary and redundant, so it was omitted.

References:

Conway, P., & Gawronski, B. (2013). Deontological and Utilitarian Inclinations in Moral Decision Making: A Process Dissociation Approach. *Journal of Personality and Social Psychology*, 104, 216-235. http://doi.org/10.1037/a0031021

杜秀芳, 刘娜娜. (2018). 金钱刺激和决策者角色对个体道德决策的影响——基于过程分离范式. 心理科学, 41(03), 667-673.

8. For experimental materials, scales, or questionnaires that have not undergone peer review, are they attached at the end of the file for review? If not, please explain why. If this article is published, are you willing to share these materials with other researchers?

Answer: Yes, please see the materials in the Appendix. If this article is published, we are willing to share these materials with other researchers.

9. This journal requires authors to provide raw data. Please choose one of the following options:

a) Send data to the editorial office email after submission

b) Data can be obtained from the following link

c) Raw data and programs have been shared on the Psychological Science Data Bank (https://psych.scidb.cn/)

d) If data cannot be provided, please explain the reason or provide relevant proof.

10. Is your study a clinical intervention or laboratory experiment? Yes (√) No ( )

If yes, please provide the pre-registration number. If no, please explain why: The first author (who executed the research) is a graduate student who has not yet developed the habit of pre-registration. Due to time constraints, pre-registration was forgotten before data collection.

Note: Clinical interventions or laboratory experiments should be pre-registered before data collection. Other experimental studies are also encouraged to pre-register. Pre-registration requires stating all research hypotheses and their rationale, plus detailed experimental/intervention procedures. This journal's pre-registration website is https://os.psych.ac.cn/preregister (see "Download Center" on the journal website for instructions) or https://osf.io/ or https://aspredicted.org/. Pre-registration significantly increases acceptance chances. For the importance of pre-registration, see https://osf.io/5awp4/

11. If your study used human or animal subjects, was it approved by your institution's ethics committee? If yes, please send a scanned copy to the editorial office email. If no, please explain.

Answer: Our study was approved by the institutional ethics committee. The ethics approval certificate has been sent to the editorial office email.

12. Did you write a 400-500 word extended English abstract following the "English Abstract Writing Guidelines" posted on the editorial office website? Has the English title and abstract been reviewed by an English-proficient professional or edited by a professional SCI/SSCI editing service?

Answer: An English abstract has been written according to the guidelines and has been edited by a professional paper polishing service.

13. If the first author is a student, please have the advisor send a separate email to the editorial office (xuebao@psych.ac.cn) stating that they have read the paper and reviewed it carefully. Have you reminded your advisor to send this email? (The editorial office will only consider the manuscript for processing after receiving the advisor's email)

Answer: The advisor has sent an email to the editorial office.

14. Please download and complete the "Manuscript Non-Confidentiality Certificate" from the "Download Center" on the right side of the editorial office website homepage, stamp it with the corresponding author's institution's confidentiality office seal, and send a scanned copy to the editorial office email (xuebao@psych.ac.cn). If there is no confidentiality office seal, please use the institution's official seal. Have you sent the email?

Answer: The "Manuscript Non-Confidentiality Certificate" has been downloaded, completed, stamped, and sent to the editorial office.

The Structure and Function of Harm Aversion Based on the Process Dissociation Procedure

Abstract

Harm aversion is central to human morality, yet its structure and mechanisms for inhibiting aggression remain unclear. Therefore, we conducted one pilot study and three formal experiments. The pilot study developed a harm action/outcome aversion dissociation task based on the process dissociation pro-

cedure. Experiment 1 used this task to examine the structure of harm aversion, while Experiment 2 (including 2a and 2b) investigated the moral cognitive mechanisms through which harm action/outcome aversion inhibits aggression. Results revealed that harm aversion comprises two independent factors: harm action aversion and harm outcome aversion. Harm outcome aversion inhibited aggression through moral disengagement, whereas the inhibitory effects of harm action aversion on aggression and moral disengagement were not robust.

These experiments developed a research tool for harm aversion, tested its two-factor structure and differential effects on inhibiting aggressive cognition and behavior, and revealed the moral cognitive pathway through which harm aversion inhibits aggression, thereby advancing theory on moral emotions and aggression.

Keywords: harm action aversion, harm outcome aversion, process dissociation procedure, aggression, moral disengagement
Classification Code: B849: C91

Harm aversion refers to the uneasy reaction or tendency to experience discomfort when individuals enact, think about, see, or hear about harmful actions or outcomes (Cushman et al., 2013; Hou et al., 2023; Miller et al., 2014). It is central to human morality (Yu et al., 2019) and serves the positive function of reducing immoral behavior while increasing prosocial behavior (Sarlo et al., 2014). Lack of harm aversion increases antisocial behavior (Crockett et al., 2014). Thus, exploring the nature and mechanisms of harm aversion holds significant scientific and practical value. However, the structure of harm aversion and its function and mechanisms for inhibiting aggression remain unclear. This study addresses these issues.

1.1 The Structure of Harm Aversion

Scholars have proposed that harm aversion is not a unitary concept but rather comprises two psychologically distinct and independent factors: harm action aversion and harm outcome aversion (Cushman et al., 2012; Miller et al., 2014; Reynolds & Conway, 2018). Harm action aversion refers to the uneasy reaction or tendency to experience discomfort when enacting harmful actions without involving negative consequences. Harm outcome aversion refers to the uneasy reaction or tendency to experience discomfort when merely encountering harmful outcomes without seeing or performing harmful actions.

One theoretical foundation for the qualitative difference between harm action/outcome aversion is the theory of dyadic morality (Gray et al., 2012). Specifically, harm action aversion reflects the perpetrator's perspective, whereas harm outcome aversion reflects the victim's perspective and involves empathy for the victim. Additionally, based on Blair's (1995) violence inhibition mechanism (VIM) and Patil's (2015) perspective, Miller et al. (2014) proposed that harm outcome aversion is innate, while harm action aversion is likely learned. In summary, harm action and outcome aversion are two qualitatively distinct psychological phenomena.

Scholars further argue that harm action and outcome aversion are independent and cannot form a higher-order harm aversion factor (Miller et al., 2014), though this claim lacks strong empirical support. For example, the only existing questionnaire based on this view—the Harm Action/Outcome Questionnaire (Miller et al., 2014)—shows high positive correlations between its two dimensions (0.54–0.66) (Miller et al., 2014; Reynolds & Conway, 2018), contradicting the hypothesis that the two factors are independent and cannot form a higher-order factor. Miller et al. (2014) explained that when completing harm action aversion items, participants unconsciously think about harm outcomes, so the subscale also measures harm outcome aversion. In other words, the developers acknowledged that the questionnaire cannot measure pure harm action aversion. Therefore, a method that can effectively separate harm action and outcome aversion is needed. The process dissociation procedure (PDP) may provide a solution.

1.2 Separating Harm Action/Outcome Aversion

Jacoby (1991) first used PDP to separate conscious and unconscious attention. Later, scholars used this paradigm to calculate utilitarian (U) and deontological (D) parameters, successfully achieving their separation (Armstrong et al., 2018; Conway & Gawronski, 2013).

Theoretically, harm action aversion and harm outcome aversion are relatively independent moral psychologies, suggesting that PDP can separate them. Methodologically, PDP separates two independent psychological components by creating incongruent conditions (where different components produce different responses) and congruent conditions (where they produce the same responses). By measuring responses in both conditions, the relative contribution of each component can be quantified, and parameters representing the two components can be calculated. These parameters should not be significantly correlated, achieving separation. Based on definitions and properties of harm action/outcome aversion, they have relatively independent effects on two types of harm-related responses (avoiding direct harmful actions and avoiding direct observation of harmful outcomes) (Cushman et al., 2012; Miller et al., 2014), meeting PDP requirements. Specifically, real-life situations exist where high harm action aversion and high harm outcome aversion individuals make the same choice (congruent conditions) or different choices (incongruent conditions).

Choices in congruent and incongruent conditions can be explained using the processing tree in Figure 1 [Figure 1: see original paper]: (1) choices dominated by harm outcome aversion (O parameter, top path); (2) choices dominated by harm action aversion (A parameter, middle path); (3) choices dominated by neither (bottom path). Additionally, "1–O" indicates situations where O parameter does not dominate, and "1–A" indicates where A parameter does not dominate. When O parameter dominates, individuals avoid harmful actions in congruent conditions but choose them in incongruent conditions. When O does not dominate but A does, individuals avoid harmful actions in both conditions. When neither dominates, individuals are more likely to choose harmful actions

in both conditions.

Figure 1. Processing tree for separating harm action aversion and harm outcome aversion (Note: A = harm action aversion parameter; O = harm outcome aversion parameter)

Thus, harm action aversion (A parameter) and harm outcome aversion (O parameter) can calculate probabilities (p) of choosing or avoiding harmful actions in congruent and incongruent conditions:

p(avoid harm|congruent) = O + [(1–O)×A]
p(choose harm|congruent) = (1–O) × (1–A)
p(avoid harm|incongruent) = (1–O) × A
p(choose harm|incongruent) = O + [(1–O) × (1–A)]

If probabilities of choosing/avoiding harm in these four situations are obtained, A and O parameters can be calculated:

O = p(avoid harm|congruent) – p(avoid harm|incongruent)
A = p(avoid harm|incongruent) / (1–O)

Thus, by completing incongruent and congruent tasks, parameters representing harm action and outcome aversion can be obtained. Theoretically, these parameters should not be significantly correlated. We propose Hypothesis 1: PDP can separate A parameter (representing harm action aversion) and O parameter (representing harm outcome aversion), and these parameters will not be significantly correlated.

1.3.1 Inhibitory Effects of Harm Action/Outcome Aversion on Aggression

Although scholars (Crockett et al., 2010; Crockett et al., 2015; Perera et al., 2016) suggest that harm aversion can inhibit aggression, empirical evidence remains insufficient, and no study has revealed the mechanisms. Therefore, this study uses harm action/outcome aversion as independent variables to reveal their inhibitory effects on aggression and underlying mediating mechanisms.

Aggression is a harmful response or tendency with intent to injure, which the victim wants to avoid (Buss & Perry, 1992). We propose that harm action aversion inhibits aggression because it represents aversion and resistance to harmful actions, while aggression is intentional harm toward others. Thus, harm action aversion should inhibit aggressive behavior (Miller et al., 2014; Miller & Cushman, 2013). We propose that harm outcome aversion inhibits aggression because it triggers avoidance responses to aggressive actions, with indirect evidence supporting this view (Blair, 1995; Buss, 1966). Accordingly, we propose Hypothesis 2: Both harm action and outcome aversion inhibit aggression.

1.3.2 Moral Cognitive Pathways of Harm Action/Outcome Aversion Inhibiting Aggression

As mentioned, harm action/outcome aversion is an important moral factor (Crockett et al., 2015; Yu et al., 2019). Therefore, moral pathways likely constitute important mechanisms. Among moral cognitive factors influencing aggression, moral disengagement is the most recognized important variable (Cen et al., 2022; Ogunfowora et al., 2022). It refers to individuals' tendency to cognitively restructure immoral actions as acceptable or moral to reduce or avoid moral inhibition and self-condemnation (Bandura, 2002; Li & Xia, 2024). As described below, harm action/outcome aversion is closely related to moral disengagement, making it a likely primary moral cognitive pathway.

We propose that harm action/outcome aversion inhibits moral disengagement because moral emotions can inhibit negative moral cognitions including moral disengagement. For example, empathy and guilt inhibit moral disengagement (Chowdhury & Fernando, 2013; Leviston & Walker, 2020). Specifically, harm action/outcome aversion involves aversion and discomfort toward harmful actions, intentions, or outcomes (Cushman et al., 2012; Miller et al., 2014), likely inhibiting cognitive distortion or beautification of harmful actions/intentions from a positive perspective. Such positive distortion of harmful actions and intentions is typical of aggression-related moral disengagement (Bjärehed et al., 2019). Thus, harm aversion likely inhibits aggression-related moral disengagement.

Moral disengagement can deactivate the moral system's regulatory function, weakening individuals' moral self-regulation (Wang et al., 2022), reducing the moral system's inhibition of aggression, and even morally endorsing harmful actions (Bjärehed et al., 2019). Therefore, moral disengagement can drive aggression (Guo et al., 2024).

We further propose that an important moral cognitive mechanism through which harm action/outcome aversion inhibits aggression is: harm action/outcome aversion inhibits individuals' moral justification or beautification of aggressive actions, thereby reducing the likelihood or frequency of aggression. Accordingly, we propose Hypothesis 3: Moral disengagement mediates the relationship between harm action/outcome aversion and aggression.

1.4 The Present Study

First, we developed a harm action/outcome aversion dissociation task based on PDP. Using experimental results, we calculated parameters representing harm action and outcome aversion, examined their validity, and tested their correlation. Finally, we used these parameters to explore the inhibitory effects of harm action/outcome aversion on aggression and the mediating role of moral disengagement.

2 Pilot Study: Development of the Harm Action/Outcome Aversion Dissociation Task

Applying PDP to separate harm action and outcome aversion requires developing tasks with congruent and incongruent conditions. The key requirement is that in congruent tasks, individuals high in both harm action and outcome aversion tend to make choices that avoid harming others. In incongruent tasks, when choosing between "seeing harm outcomes without directly enacting harm" versus "enacting harm without directly seeing outcomes," high harm action aversion individuals prefer the former, while high harm outcome aversion individuals prefer the latter. We adapted moral dilemma tasks (Armstrong et al., 2018; Conway & Gawronski, 2013) to develop a task meeting these requirements.

2.1.1 Participants

We recruited 42 university students offline (from our university) and via an online survey platform (22 and 20 respectively). After excluding 2 inattentive participants, we obtained 40 valid participants (19 male, 21 female), aged 18.08–26.83 years (M = 21.52, SD = 2.68).

2.1.2 Materials and Procedure

The harm action/outcome aversion dissociation task used scenario materials reflecting incongruent and congruent conditions. Initial 32 scenarios were developed by a psychology professor and a psychology student based on relevant literature (Conway & Gawronski, 2013; Miller et al., 2014; Patil, 2015). Six psychology graduate and undergraduate students rated each scenario for choice difficulty, imagination difficulty, and provided modification suggestions. Based on ratings and suggestions, we discussed, screened, and revised scenarios multiple times, finally creating 10 scenarios: "Taken Hostage," "Ordered to Torture a Stranger," "Choosing a Transmigrated Identity," "Completing a Spy Mission," "Battlefield Rescue," "Script Selection for Performance," "Defending a Friend," "Retaliating Against a Scammer," "Undercover Self-Preservation," and "Dealing with a Competitor."

Each scenario included three parts: a situational story, options for incongruent conditions, and options for congruent conditions. For example, in the "Dealing with a Competitor" scenario, the story was: "Suppose you have a competitor in class who frequently provokes you, making you feel bad. What would you do?" Incongruent options were: (1) "After being provoked, you and friends beat a punching bag mannequin in a venting room as if it were him/her, cursing while hitting" (Note: This represents "enacting harm without directly seeing outcomes"; high harm action aversion individuals typically avoid this, while high harm outcome aversion individuals typically choose it) and (2) "You don't get involved; you let friends defend you, and for many days afterward you see him/her come to class bruised and dispirited" (Note: This represents "seeing harm outcomes without directly enacting harm"; high harm action aversion individuals

typically choose this, while high harm outcome aversion individuals typically avoid it). Congruent options were: (1) "With several friends, you find his/her social media account and anonymously insult him/her in the comments; for a week you see him/her become depressed and dispirited as a result" (Note: This represents "enacting harm while seeing outcomes"; high harm action/outcome aversion individuals typically avoid this) and (2) "You spend all your allowance on tutoring classes, studying hard over 12 hours daily to distract yourself" (Note: This represents "non-harmful action"; high harm action/outcome aversion individuals both choose this).

Participants first made choices for each of the 10 scenarios under both incongruent and congruent conditions, then rated choice difficulty and imagination difficulty for each scenario, and provided modification suggestions. Both difficulty ratings used 7-point Likert scales (1 = very easy, 4 = moderate, 7 = very difficult).

Results showed that participants chose harmful actions in 38.75% of congruent trials and 57.50% of incongruent trials. This pattern resembles Conway and Gawronski's (2013) PDP results separating deontology and utilitarianism (28% accepted harm in congruent, 58% in incongruent conditions). The results also align with the hypothesis that high harm action/outcome aversion individuals make fewer harmful choices in congruent conditions.

Using formulas 5 and 6 from the Introduction, we calculated A and O parameters representing harm action and outcome aversion. Correlation analysis revealed these parameters were not significantly correlated ($r = 0.22$, $p = 0.182$), consistent with the view that parameters separating utilitarianism and deontology should be uncorrelated or weakly correlated (Armstrong et al., 2018; Conway & Gawronski, 2013).

Additionally, choice difficulty ratings did not differ significantly between congruent ($M = 3.48$, $SD = 1.15$) and incongruent ($M = 3.27$, $SD = 1.16$) conditions, $t(39) = 1.80$, $p = 0.080$, Cohen's $d = 0.28$. Scenario imagination difficulty ratings fell between "relatively easy" and "moderate" ($M = 3.78$, $SD = 1.80$). This suggests comparable difficulty between incongruent and congruent conditions, and that participants could adequately imagine and immerse themselves in the scenarios.

Based on completion rates, ratings, and suggestions, we further discussed and optimized each scenario's wording, finalizing the formal harm action/outcome aversion dissociation task.

Overall, pilot results suggest that the PDP-based task can obtain relatively independent harm action and outcome aversion parameters, effectively separating the two constructs and providing preliminary support for Hypothesis 1.

3 Experiment 1: Testing the Two-Factor Structure of Harm Aversion

Experiment 1 tested the hypothesis that harm aversion comprises two independent factors. We examined relationships between A and O parameters and their validity using criterion variables.

We measured harm action/outcome aversion using the Harm Action/Outcome Questionnaire, along with criterion variables including empathy, moral judgment, deontology D parameter, psychopathy, physical aggression, and verbal aggression. First, empathy positively correlates with both harm action and outcome aversion, with a stronger relationship to harm outcome aversion (Miller et al., 2014). Second, both harm action and outcome aversion play important roles in moral judgment (Conway & Gawronski, 2013; Wiech et al., 2013). Third, both correlate positively with the D parameter (Reynolds & Conway, 2018). Fourth, psychopathy negatively correlates with both, with a stronger negative relationship to harm outcome aversion (Patil, 2015; Reynolds & Conway, 2018). Fifth, harm aversion can inhibit aggression (Blair, 1995, 2007; Crockett et al., 2015; Cushman et al., 2013). Additionally, because the Harm Action/Outcome Questionnaire's action subscale involves harm outcomes, its scores likely correlate significantly with the O parameter.

Thus, Experiment 1 had four hypotheses: Hypothesis 1a: A and O parameters will show low, non-significant correlation; Hypothesis 1b: A parameter will correlate positively with harm action questionnaire scores, while O parameter will correlate positively with both harm outcome and action questionnaire scores; Hypothesis 1c: Both parameters will correlate positively with empathy, moral judgment, and D parameter, and negatively with psychopathy and aggression; Hypothesis 1d: Compared to A parameter, O parameter will show stronger positive correlation with empathy and stronger negative correlation with psychopathy.

3.1 Participants

Using G*Power with effect size r = 0.2, two-tailed test ($\alpha = 0.05$), 218 participants were needed for 0.85 power. We recruited 347 university students from three universities, excluding 6 participants with completion times under 5 minutes or inattentive responses. Following PDP rules (Conway & Gawronski, 2013), we deleted 54 participants who consistently avoided harm in congruent conditions. The final sample included 287 participants (103 male, 184 female), aged 18.08–23.75 years (M = 20.32, SD = 1.30). Each received 6 RMB compensation.

3.2.1 Harm Action/Outcome Aversion Dissociation Task

We used the pilot-developed task to obtain harm action and outcome aversion parameters. The task included 10 scenarios, each with incongruent and congruent conditions and two options per condition.

3.2.2 Harm Action/Outcome Aversion Questionnaire

We used Miller et al.'s (2014) Harm Action/Outcome (A/Q) Questionnaire, which has 34 items: 9 measuring harm action aversion (e.g., "In a performance, you stab an actor's neck with a retractable prop knife"), 14 measuring harm outcome aversion (e.g., "You see someone close a car door on their finger"), and 11 control items to avoid response bias and conceal the survey's purpose (e.g., "You listen to 'Happy Birthday' 100 times in a row"). Items were rated 1 (not at all) to 7 (very strongly) for discomfort experienced, with higher scores indicating greater harm aversion. The English (Miller et al., 2014; Patil, 2015) and Chinese (Hou et al., 2023) versions show good reliability and validity. In this study, the questionnaire demonstrated good structural validity: ^{2}/df = 49.83/19, RMSEA [90% CI] = 0.07[0.050, 0.101], CFI = 0.98, TLI = 0.96, SRMR = 0.03. Cronbach's α was 0.88 for harm action aversion and 0.92 for harm outcome aversion.

3.2.3 Empathy Subscale

We used the Empathic Concern subscale from Davis's (1983) Interpersonal Reactivity Index (IRI), which includes 7 items (e.g., "I often have tender, concerned feelings for people less fortunate than me"). Items were rated 1 (does not describe me well) to 5 (describes me very well). The English (Davis, 1983) and Chinese (Rong et al., 2010) versions show good reliability and validity. In this study, Cronbach's α was 0.67.

3.2.4 Moral Dilemma Task

We used Conway and Gawronski's (2013) moral dilemma task to obtain the deontology D parameter. The task includes 10 moral dilemma scenarios, each with two versions for incongruent and congruent conditions. In both conditions, participants judged whether a harmful action was acceptable. Incongruent conditions involved harmful actions yielding large benefits, while congruent conditions involved actions yielding small benefits. High deontologists tend to reject harmful actions in incongruent conditions, while utilitarians tend to accept them. In congruent conditions, both high utilitarians and deontologists tend to reject harmful actions.

3.2.5 Levenson Psychopathy Scale

We used the Levenson Self-Report Psychopathy Scale (Levenson et al., 1995) to measure psychopathic traits (e.g., "In today's world, I feel that anything is justified to achieve success"). Items were rated 1 (strongly disagree) to 4 (strongly agree). The English (Levenson et al., 1995) and Chinese (Deng et al., 2017) versions show good reliability and validity. In this study, Cronbach's α was 0.80.

3.2.6 Physical and Verbal Aggression Subscales

We used the Physical Aggression (9 items; e.g., "If provoked enough, I may hit someone") and Verbal Aggression (5 items; e.g., "When people annoy me, I may tell them what I think of them") subscales from the Buss-Perry Aggression Questionnaire (BPAQ; Buss & Perry, 1992). Items were rated 1 (extremely uncharacteristic of me) to 5 (extremely characteristic of me). The English (Buss & Perry, 1992) and Chinese (Quan et al., 2019) versions show good reliability and validity. In this study, Cronbach's α was 0.82 for physical aggression and 0.72 for verbal aggression.

3.3 Procedure

Participants completed the harm action/outcome aversion dissociation task, harm action/outcome aversion questionnaire, IRI empathic concern subscale, moral dilemma task, Levenson Psychopathy Scale, BPAQ physical and verbal aggression subscales, and a demographic questionnaire offline.

3.4 Statistical Analysis

We used SPSS 24.0 for analyses. First, we analyzed choice proportions in congruent and incongruent conditions. Second, we calculated A and O parameters using Silver's Z program in R. Third, we computed correlations between A/O parameters and criterion variables. Finally, we conducted significance tests for differences between correlation coefficients (Diedenhofen & Musch, 2015).

3.5.1 Harm Action Choice Proportion Analysis

Participants chose harmful actions in 62.40% (SD = 0.18) of incongruent trials and 27.98% (SD = 0.18) of congruent trials. These proportions resemble Conway and Gawronski's (2013) findings (28% accepted harm in congruent, 58% in incongruent conditions). A paired-samples t-test showed significantly higher harm choices in incongruent versus congruent conditions, $t(286) = 20.80$, $p < 0.001$.

3.5.2 Parameter Validity Tests

Table 1 Correlations between A/O parameters and criterion variables

Criterion Variable	A Parameter	O Parameter	z-test for difference	p-value
Harm Action Aversion (Questionnaire)	0.16**	0.13*	-2.02*	0.044
Harm Outcome Aversion (Questionnaire)	0.27***	0.23***	-2.25*	0.024
Traditional Moral Judgment	0.23***	0.30***	0.23***	-
Deontology D Parameter	0.29***	0.28***	-0.13*	0.894
Utilitarianism U Parameter	-0.27***	-0.22***	-0.13*	0.894
Levenson Psychopathy	-0.22***	-0.42***	2.07*	0.038
Physical Aggression	-0.22***	-0.22***	-	-

Criterion Variable	A Parameter	O Parameter	z-test for difference	p-value
Verbal Aggression	-0.13*	-0.22***	-	-

Note: A parameter = harm action aversion parameter; O parameter = harm outcome aversion parameter; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Correlation analyses showed A and O parameters were not significantly correlated ($r = 0.10$, $p = 0.106$), suggesting the two psychological components are relatively independent, supporting Hypothesis 1a and scholars' views (Cushman, 2013; Cushman et al., 2012; Miller et al., 2014).

Results supported that A and O parameters represent harm action and outcome aversion. First, harm action questionnaire scores correlated positively with A parameter, while harm outcome questionnaire scores correlated positively with O parameter but not A parameter. The positive correlation between harm action questionnaire scores and O parameter supported Miller et al.'s (2014) claim that their harm action subscale involves harm outcome aversion. Second, empathic concern correlated positively with both parameters, with a significantly stronger correlation with O parameter ($z = –2.25$, $p = 0.024$), consistent with findings that empathy relates more closely to harm outcome aversion (Cushman et al., 2012; Miller et al., 2014). Third, psychopathy correlated negatively with both parameters, with a significantly stronger negative correlation with O parameter ($z = 2.07$, $p = 0.038$), consistent with stronger negative relationships between harm outcome aversion and psychopathy (Patil, 2015; Reynolds & Conway, 2018). Fourth, moral judgment correlated positively with both parameters, aligning with research showing positive correlations between moral judgment and harm action/outcome aversion (Reynolds & Conway, 2018; Wiech et al., 2013). Fifth, deontology D parameter correlated positively with both parameters, consistent with findings that deontology parameters correlate positively with harm action/outcome aversion (Cushman, 2013; Reynolds & Conway, 2018). Utilitarianism U parameter did not correlate significantly with either parameter. Sixth, physical aggression correlated negatively with both parameters; verbal aggression correlated significantly only with O parameter.

Some unexpected results emerged. Correlations between A/O parameters and questionnaire scores were modest (0.16–0.23), possibly because the Harm Action/Outcome Questionnaire (Miller et al., 2014) focuses on non-violent harm, while our dissociation task emphasizes violent harm. Violent harm is the primary form of harm and the main source of harm aversion, so measurement should be based on violent harm. Additionally, A parameter did not correlate significantly with verbal aggression, likely because verbal aggression causes relatively low and subjective harm (only harmful when the victim cares). This suggests harm action aversion's inhibitory function may differ across aggression types.

4 Experiment 2: Moral Cognitive Mechanisms of Harm Action/Outcome Aversion Inhibiting Aggression

Although scholars generally agree that harm aversion inhibits aggression, empirical evidence is insufficient and mechanisms remain unclear. Experiment 2 explored the inhibitory effects of harm action/outcome aversion on aggression and the mediating role of moral disengagement.

Experiment 1 showed differential relationships between harm action/outcome aversion and verbal/physical aggression. Therefore, Experiment 2a retained verbal and physical aggression as dependent variables to further examine these relationships while investigating mediating pathways.

4.1.1 Participants

Using G*Power for regression with 5 predictors (A parameter, O parameter, moral disengagement, gender, age) and effect size $f^2 = 0.07$, $\alpha = 0.05$, 212 participants were needed for 0.85 power. With two potential dependent variables, we required 424 participants.

We recruited 518 university students from four universities, excluding 17 inattentive participants and 78 who consistently avoided harm in congruent conditions per PDP rules. The final sample included 423 participants (170 male, 253 female), aged 18.08–23.75 years (M = 20.32, SD = 1.30), who received 6 RMB compensation.

4.1.2 Procedure

Due to COVID-19, participants completed the harm action/outcome aversion dissociation task, Civic Moral Disengagement Questionnaire, and BPAQ physical and verbal aggression subscales through combined online and offline methods. Average completion time was 10 minutes.

4.1.3 Measures

Harm Action/Outcome Aversion Dissociation Task: Same as Experiment 1.

Civic Moral Disengagement Questionnaire: We used Caprara et al.'s (1996) Civic Moral Disengagement Questionnaire, which has 32 items (e.g., "To protect one's interests, using force is often inevitable"). Items were rated 1 (strongly disagree) to 5 (strongly agree). The scale shows good reliability and validity in previous research (Caprara et al., 2009; Wang et al., 2013). In this study, Cronbach's α was 0.928.

Physical and Verbal Aggression Subscales: Same as Experiment 1, with good reliability (verbal aggression Cronbach's $\alpha = 0.75$; physical aggression Cronbach's $\alpha = 0.84$).

4.1.4 Statistical Analysis

We used SPSS 24.0 and Mplus 8.0. First, we conducted descriptive statistics and correlations. Second, we parcelled items and performed confirmatory factor analysis using maximum likelihood estimation for missing and non-normal data. Finally, we conducted structural equation modeling and tested mediation using bias-corrected Bootstrap with 5,000 resamples. Good model fit was indicated by: ^{2}/df < 5, RMSEA and SRMR < 0.08, TLI and CFI > 0.90 (Hoyle & Panter, 1995).

4.1.5 Results and Discussion

Descriptive Statistics and Correlations: As shown in Table 2 , A and O parameters were not significantly correlated. Both parameters correlated negatively with moral disengagement and physical aggression, while moral disengagement, physical aggression, and verbal aggression correlated positively with each other. Verbal aggression correlated significantly only with O parameter, not A parameter. Therefore, we did not examine mediation of moral disengagement between A parameter and verbal aggression.

Table 2 Descriptive statistics and correlations

Variable	1	2	3	4	5
1. A Parameter	1				
2. O Parameter	0.08	1			
3. Moral Disengagement	-0.17***	-0.32***	1		
4. Physical Aggression	-0.16**	-0.32***	0.60***	1	
5. Verbal Aggression	-0.13**	-0.22***	0.48***	0.32***	1

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

Mediation Analysis: To reveal mechanisms, we tested whether moral disengagement mediated relationships between A/O parameters and aggression (Figure 2 [Figure 2: see original paper]), controlling for gender and age. The mediation model showed good fit: ^{2}/df = 229.13/87, RMSEA [90% CI] = 0.06[0.052, 0.072], CFI = 0.94, TLI = 0.93, SRMR = 0.05. A parameter negatively predicted physical aggression (β = –0.12, p = 0.020, 95% CI = [–0.233, –0.021]) but not verbal aggression (β = –0.04, p = 0.525, 95% CI = [–0.147, 0.073]). O parameter negatively predicted both physical (β = –0.26, $p < 0.001$, 95% CI = [–0.383, –0.120]) and verbal aggression (β = –0.14, p = 0.041, 95% CI = [–0.267, –0.006]). These results support the view that harm aversion inhibits aggression (Blair, 1995; Crockett et al., 2015), consistent with findings that harm aversion reduces harmful behavior (Crockett et al., 2010; Perera et al., 2016).

Bootstrap indirect effect tests revealed that moral disengagement mediated between A parameter and physical aggression (β = –0.08, p = 0.010, 95% CI =

[–0.272, –0.098]) and between O parameter and physical (β = –0.17, p < 0.001, 95% CI = [–0.272, –0.098]) and verbal aggression (β = –0.10, p = 0.002, 95% CI = [–0.179, –0.049]). Direct effects of A parameter on physical (β = –0.04, p = 0.343, 95% CI = [–0.195, 0.020]) and verbal aggression (β = 0.01, p = 0.821, 95% CI = [–0.167, 0.093]) were non-significant, as were direct effects of O parameter on physical (β = –0.09, p = 0.110, 95% CI = [–0.195, 0.020]) and verbal aggression (β = –0.04, p = 0.579, 95% CI = [–0.167, 0.093]). These results align with findings that moral emotions negatively predict moral disengagement (Chowdhury & Fernando, 2013; Ouvrein et al., 2018) and that moral disengagement positively predicts aggression (Gini et al., 2021; Teng et al., 2019).

Figure 2 [**Figure 2: see original paper**] Structural model of A/O parameters, moral disengagement, and physical/verbal aggression

Note: Path coefficients are standardized. *$p < 0.05$;* ***$p < 0.01$;*** p < 0.001. Paths for gender and age are hidden for clarity. Dashed lines indicate non-significant paths.

Experiment 2a used questionnaires to measure aggression, which cannot effectively test causality. Experiment 2b further tested mediation using experimental methods.

4.2.1 Participants

Using G*Power with effect size f^2 = 0.15 and 5 predictors (A parameter, O parameter, moral disengagement, gender, age), 102 participants were needed for 0.85 power. We recruited 124 participants two weeks before the formal experiment and had them complete the harm action/outcome aversion dissociation task. Following calculation rules, we excluded participants who consistently avoided harm in congruent conditions or responded inattentively, leaving 107 valid participants (33 male, 74 female), aged 17.83–28.17 years (M = 21.12, SD = 1.73). All were right-handed with normal/corrected vision, no psychiatric history, and willing to undergo electric shock procedures. Participants received 20–35 RMB compensation based on choices and attentiveness.

4.2.2 Materials

Harm Action/Outcome Aversion Dissociation Task: Same as Experiments 1 and 2a.

Adapted Pain-Gain Task: The pain-gain task (PGT; Feldmanhall et al., 2015) gives participants ("deciders") money and asks how much they would sacrifice to prevent a "receiver" from receiving electric shocks—more money sacrificed means fewer painful shocks for the receiver. We adapted this to measure aggression and moral disengagement. First, we modified the task so participants earned money by choosing to shock the receiver, turning it into an aggression task. Second, we added a "moral appropriateness" rating before shock choices to measure moral disengagement.

Electric Stimulator: We used a commercial electric stimulator (YRKJ-F1002) to deliver 2-second transcutaneous electric shocks via two 1-cm AgCl electrodes placed ~2 cm apart on participants' left forearm. Before the experiment, we calibrated individual pain thresholds. During and after the experiment, participants believed they were shocking another person (though no actual shocks occurred).

4.2.3 Procedure

Upon arrival: (1) Participants read instructions and provided informed consent. They were told there were two roles—"decider" and "receiver"—and that as deciders, they could shock receivers for extra money. The matched receiver was described as a same-sex student in another room completing a memory task under shock risk.

(2) We measured pain thresholds using the stimulator and a 9-point scale ("pain that is unbearable upon contact"), calibrating all participants and informing them that receivers would receive equally painful shocks.

(3) Participants completed two practice rounds, with shocks and money delivered after each round. If they chose to shock, they pressed the shock button themselves. They were told a shock cable connected to another room where participants wouldn't meet (the cable was fake; a confederate sat silently in the adjacent room).

(4) Participants completed the formal experiment: 40 critical trials and 10 filler trials, divided into two blocks of 25 trials each (20 critical, 5 filler) in random order. Critical trials varied shock duration (2–3 seconds) and money amounts (3–5 RMB), randomly paired. Filler trials with high money/low shock or low money/high shock prevented response sets and experimental purpose guessing.

The trial procedure (Figure 3 [Figure 3: see original paper]) was: fixation cross "+" → money and shock duration cue → participants considered whether to shock and rated moral appropriateness (1 = very inappropriate, 4 = very appropriate) → fixation cross → participants pressed keys to decide ("~" = shock, "/" = no shock; positions randomized) → feedback on money earned.

After the experiment, participants were interviewed and compensated based on selected trials.

Figure 3 [Figure 3: see original paper] Example trial in the adapted pain-gain task

4.2.4 Statistical Analysis

We used SPSS 24.0 for descriptive statistics and correlations, then the Process plugin with 5,000 Bootstrap resamples to test indirect effects of moral disengagement between A/O parameters and aggressive behavior.

4.2.5 Results and Discussion

Descriptive Statistics and Correlations: Table 3 shows: (1) A and O parameters were not significantly correlated (r = 0.17, p = 0.087), consistent with Experiments 1 and 2a; (2) O parameter correlated negatively with moral disengagement (r = –0.21, p = 0.028) and aggressive behavior (r = –0.24, p = 0.014), consistent with Experiment 2a; (3) A parameter did not correlate significantly with moral disengagement (r = –0.02, p = 0.854) or aggressive behavior (r = –0.01, p = 0.963), inconsistent with Experiment 2a. Therefore, only O parameter was included in subsequent mediation analysis.

Table 3 Descriptive statistics and correlations

Variable	1	2	3	4
1. A Parameter	1			
2. O Parameter	0.17	1		
3. Moral Disengagement	-0.02	-0.21*	1	
4. Aggressive Behavior	-0.01	-0.24*	0.60***	1

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

Mediation Analysis: As shown in Figure 4 [Figure 4: see original paper], O parameter significantly negatively predicted moral disengagement (β = –0.21, p = 0.008), which positively predicted aggressive behavior (β = 0.57, $p < 0.001$). O parameter's direct effect on aggressive behavior was non-significant (β = –0.12, p = 0.117, 95% CI = [–0.261, 0.033]), while the indirect effect through moral disengagement was significant (β = –0.12, p = 0.013, 95% CI = [–0.209, –0.024]), consistent with Experiment 2a.

Figure 4 [Figure 4: see original paper] Mediating role of moral disengagement between O parameter and aggressive behavior

5 General Discussion

To explore the structure of harm aversion and its effects and pathways for inhibiting aggression, we conducted one pilot study and three formal experiments, developing a harm action/outcome aversion dissociation task. Key findings: (1) A parameter (harm action aversion) and O parameter (harm outcome aversion) were not significantly correlated, yet both correlated positively with harm action aversion questionnaire scores, empathy, moral judgment, and deontology parameter, and negatively with psychopathy and physical aggression. O parameter showed stronger relationships with empathy, psychopathy, and verbal aggression than A parameter. A parameter did not correlate significantly with harm outcome aversion questionnaire scores. (2) Harm outcome aversion negatively predicted aggression through moral disengagement mediation. (3) In Experiment 2a, harm action aversion negatively predicted physical aggression through moral disengagement, but this mediation was non-significant in Experiment 2b.

5.1 The Structure of Harm Aversion

Based on PDP, the pilot study developed a harm action/outcome aversion dissociation task. Experiment 1 showed that A and O parameters had low, non-significant correlation. Additionally, A parameter did not correlate significantly with harm outcome aversion questionnaire scores. These results support the view that harm action and outcome aversion are independent (Cushman et al., 2012; Cushman et al., 2013; Miller et al., 2014).

Experiment 1 found that O parameter showed stronger relationships with empathic concern, psychopathy, and verbal aggression than A parameter. These results align with questionnaire-based studies (Miller et al., 2014; Patil, 2015; Reynolds & Conway, 2018), suggesting qualitative and functional differences between harm action and outcome aversion.

Experiment 1 also showed that O parameter correlated with both harm action and outcome aversion questionnaire scores, supporting Miller et al.'s (2014) speculation that their harm action subscale involves harm outcome aversion.

In summary, Experiment 1 supports the hypothesis that harm aversion comprises two independent factors—harm action and outcome aversion—that differ qualitatively and functionally.

5.2 Moral Cognitive Mechanisms of Harm Aversion Inhibiting Aggression

Both Experiments 2a and 2b found that moral disengagement mediated between harm outcome aversion and aggression, suggesting harm outcome aversion inhibits aggression through this moral cognitive pathway.

Experiments 1 and 2 (including 2a and 2b) showed that harm outcome aversion correlated negatively with both physical and verbal aggression and inhibited aggressive behavior. These results align with theoretical views and indirect evidence. Many scholars argue that harm aversion inhibits aggression (Blair, 1995, 2007; Cushman, 2013). Indirect evidence includes findings that hearing victims' pain reduces shock delivery (Buss, 1966), serotonin promotes harm aversion and inhibits harm toward other players (Crockett et al., 2010), and more harm aversion in "taking" versus "receiving" conditions leads to less harm (Perera et al., 2016). Compared to previous research, Experiment 2 provides clear, direct empirical evidence that harm outcome aversion inhibits aggression, advancing research on the harm aversion-aggression relationship. This likely occurred because aggressive behavior readily leads to negative consequences, and aversion to these consequences triggers strong moral inhibition.

Both sub-studies found that harm outcome aversion negatively predicted moral disengagement, consistent with research on moral emotions and moral disengagement. For example, empathy and guilt inhibit moral disengagement (Chowdhury & Fernando, 2013; Leviston & Walker, 2020; Ouvrein et al., 2018). This likely occurs because aggressive actions typically cause harm outcomes, and

aversion to these outcomes makes it difficult to positively distort or beautify aggressive actions/intentions, which is typical of aggression-related moral disengagement (Bjärehed et al., 2019).

Both sub-studies showed positive relationships between moral disengagement and aggression, matching extensive research showing moral disengagement promotes various types of aggression (Gini et al., 2021; Nocera et al., 2022; Teng et al., 2019). This occurs because while moral systems inhibit aggression by deeming it wrong (Bandura, 2002; Fitouchi et al., 2022; Hou et al., 2023), moral disengagement can weaken or deactivate this inhibition by morally rationalizing aggression (Bjärehed et al., 2019; Wang et al., 2022), thereby promoting aggression (Guo et al., 2024).

Both sub-studies found that moral disengagement mediated between harm outcome aversion and aggression, consistent with findings that moral disengagement mediates between moral emotions (e.g., empathy, guilt) and aggression (Ouvrein et al., 2018; Wang et al., 2017). These results suggest moral disengagement is an important moral cognitive mechanism through which moral emotions inhibit aggression. Harm outcome aversion makes it difficult to morally beautify actions causing harm (like aggression), thereby inhibiting aggression.

5.3 Inhibitory Effect of Harm Action Aversion on Aggression

Experiments 1 and 2a found negative correlations between harm action aversion and physical aggression, but non-significant correlations with verbal aggression. Experiment 2b found non-significant inhibitory effects of harm action aversion on aggressive behavior. These results suggest the relationship between harm action aversion and aggression is not robust. Two possible explanations exist. First, harm action aversion (A parameter) emphasizes aversion to harmful actions themselves, excluding harm outcomes. Without harm outcomes, the moral inhibition triggered by this moral emotion may be weak, resulting in weak inhibition of aggression. Second, the nature of harmful actions likely affects the strength of moral inhibition triggered by harm action aversion. For example, research comparing trolley and footbridge dilemmas found that few people chose to push one person to save many in the footbridge dilemma, while more chose to flip a switch in the trolley dilemma (Cushman et al., 2006; Greene et al., 2009). "Pushing" is a typical, direct, intense harmful action, while "flipping a switch" is atypical (Cushman et al., 2010; Miller et al., 2014). Thus, harm action aversion likely triggered stronger moral inhibition in the footbridge dilemma, suppressing harmful choices. Compared to physical aggression, verbal aggression is a less typical harmful action, so harm action aversion may trigger weaker moral inhibition, making its correlation with verbal aggression often non-significant. Similarly, keypress shock selection in Experiment 2b is atypical, indirect physical aggression, and only some trials were actually delivered, which may explain why harm action aversion's relationship with shock selection was non-significant. These speculations require further testing.

Experiment 2a showed a significant negative correlation between harm action aversion and moral disengagement, but Experiment 2b did not, suggesting this relationship is also not robust. Two possible reasons exist. First, harm action aversion itself may poorly inhibit positive cognitive distortion of harmful actions. Without harm outcomes, individuals may not view harmful actions as highly immoral, making moral justification easier. Second, the relationship may be influenced by methodology. Experiment 2a used questionnaires to measure moral disengagement, while Experiment 2b used an experimental task. The keypress selection in Experiment 2b is atypical, and participants saw no direct harm outcomes, likely weakening moral inhibition of rationalizing shock choices. These hypotheses also require future testing.

5.4 Contributions and Limitations

This study's methods, results, and perspectives have theoretical and practical value. We tested the two-factor structure of harm aversion, revealed differential effects of harm action and outcome aversion in inhibiting aggressive cognition and behavior, and identified an important moral cognitive mechanism. These findings advance harm aversion and aggression theory, provide a theoretical foundation for future research, and have implications for aggression prevention/intervention and moral education.

This study demonstrates that PDP can measure moral emotions, and our dissociation task can be used in future harm aversion research to reveal functional and mechanistic differences. Results suggest that simultaneously enhancing both harm action and outcome aversion may better inhibit aggressive cognition and behavior and improve moral character.

Limitations include: First, our dissociation task involves more physical aggression scenarios than verbal or indirect aggression scenarios, potentially underestimating relationships with verbal/indirect aggression. Future research should develop tasks involving multiple aggression types. Second, inconsistent results for harm action aversion's relationships with moral disengagement and aggression require further investigation, particularly examining boundary conditions for these relationships.

Appendix 1: Harm Action/Outcome Aversion Dissociation Task

Incongruent Condition

1-1: You're kidnapped by a deranged masked criminal to a wooden house. You discover he's also kidnapped others and is torturing them. He tells you that to survive, you must choose:

- Option 1: Whip another blindfolded hostage until they scream in pain
- Option 2: He takes all your savings and valuables

1-2: You're kidnapped to a secret room. You discover he's also kidnapped others and is torturing and intimidating them. He tells you that to survive, you must choose:
- Option 1: Hammer 10 nails into a very realistic dummy to intimidate others
- Option 2: Watch a blood-covered hostage scream in pain for over 20 minutes

2-1: Imagine you're a member of a secret organization in the Middle East. You receive orders to torture a stranger into confessing and serving 1 year in prison. Assuming only two methods exist, you choose:
- Option 1: Personally torture the stranger with various instruments until they confess under duress
- Option 2: You disobey orders and are placed in solitary confinement, living alone in filth for half a month

2-2: Imagine you're a member of a secret organization in the Middle East. You receive orders to torture a stranger into confessing and serving 1 year in prison. Assuming only two methods exist, you choose:
- Option 1: You personally torture the stranger at low intensity for days combined with psychological tactics until they confess without real injury
- Option 2: You have subordinates do the torturing; when they signal readiness, you enter and watch the tortured, injured stranger confess in agony

[Additional scenarios 3-10 follow the same format, presenting incongruent and congruent conditions for each situation]

Congruent Condition
[Corresponding congruent versions for each scenario are presented, where both high harm action and outcome aversion individuals would make the same non-harmful choice]

Appendix 2: Harm Action/Outcome Aversion Questionnaire Items

Harm Action Aversion Items:
1. You point an empty gun at your friend's head and pull the trigger (you both know it's empty)
2. You make obscene gestures behind your best friend's back
3. In a performance, you stab an actor's neck with a retractable prop knife
4. In a realistic but safe magic show, you volunteer to "saw a woman in half"
5. With your friend's permission, you fire a gun at them behind bulletproof glass

Harm Outcome Aversion Items:
21. You see a woman in sandals accidentally run over her own toes with a shopping cart
[Additional items continue…]

Appendix 3: Moral Dilemma Task

Incongruent Condition:
You find a time machine and travel back to 1920. At a hotel, you meet a young man. You know this is George Blackman, who later kidnaps a child for a week until ransom is paid. He's in the next room with an unlocked door. It's easy to smother him with a pillow while he sleeps without being discovered, preventing the kidnapping, though he hasn't committed any crime yet.
Is it appropriate to kill George to prevent him from kidnapping a child?

Congruent Condition:
[Corresponding scenario with lower stakes where both deontologists and utilitarians would reject the harmful action]

[Additional dilemmas follow the same structure]

Appendix 4: Empathy Subscale

Items:
1. I often have tender, concerned feelings for people less fortunate than me
2. Sometimes I don't feel very sorry for other people when they are having problems (R)
3. When I see someone being taken advantage of, I feel kind of protective toward them
4. Other people's misfortunes do not usually disturb me a great deal (R)
5. When I see someone being treated unfairly, I sometimes don't feel very much pity for them (R)

Appendix 5: Levenson Psychopathy Scale

Items:
1. Success is based on survival of the fittest; I don't care about the losers
2. For me, what's right is whatever I can get away with
3. In today's world, I feel that anything is justified to achieve success
4. My main purpose in life is getting as many goodies as I can
5. I let others worry about high-minded principles; I just look out for myself
[Additional items continue…]

Appendix 6: Civic Moral Disengagement Questionnaire

Items:

1. Since there aren't adequate waste disposal facilities, it's not fair to blame people who throw trash in the street
2. It's okay to forget to fill out invoice information since checking for such errors is the finance staff's responsibility
3. Since others have committed more serious vandalism, there's no reason to punish those who only graffiti walls
4. When all cars on the road are speeding, drivers who speed to keep up shouldn't be penalized
5. Since society as a whole causes environmental degradation, individual concern about the problem is meaningless

[Additional items continue…]

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

Source: ChinaXiv — Machine translation. Verify with original.