Journal of Experimental Social Psychology 112 (2024) 104597

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Moral violations that target more valued victims elicit more anger, but not necessarily more disgust☆

Lei Fan [a,b,*], Catherine Molho [a,b], Tom R. Kupfer [c], Joshua M. Tybur [a,b]

[a] Vrije Universiteit Amsterdam, Department of Experimental and Applied Psychology, Van der Boechorststraat 7, 1081 BT Amsterdam, the Netherlands
[b] Institute for Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[c] Nottingham Trent University, Department of Psychology, England, United Kingdom

ARTICLE INFO

Keywords:
Interpersonal value
Morality
Disgust
Anger
Aggression

ABSTRACT

The same moral violation can give rise to different emotional and behavioral responses in different individuals. The mechanisms that give rise to such differences – and the functions that those mechanisms serve – are unclear. Previous work suggests that people experience greater anger toward violations that target themselves or kin than those that target others, whereas they experience greater disgust toward violations that target others than those that target themselves or kin. In turn, anger has a stronger relation with direct aggression than indirect aggression, and disgust a stronger relation with indirect aggression than direct aggression. The current study tests whether these patterns depend on the value observers place on the targets of moral violations, even within folk relationship categories. In two studies, we asked participants to think of a person they know and to imagine that person being targeted by a moral violation described in a vignette. We assessed the value that participants placed on the target using a financial tradeoff task, their emotional reaction to the violation, and their desires to aggress toward the perpetrator. Results revealed that: (1) interpersonal value relates more strongly to anger than disgust toward the moral violation; (2) interpersonal value relates more strongly to direct than indirect aggression motives; and (3) anger relates to both direct and indirect aggression motives, whereas disgust relates only to indirect aggression motives. These results suggest that the value one places on the victims of moral violations influences emotional and behavioral reactions to those violations.

1. Introduction

People often respond to moral violations with outrage, which is primarily characterized by two emotions: anger and disgust (Fan, Molho, Kupfer, Sauter, & Tybur, 2023; Rozin, Lowery, Imada, & Haidt, 1999; Salerno & Peter-Hagene, 2013). Some perspectives suggest that differences in emotional responses to moral violations are caused by differences in the content of those violations (e.g., anger for autonomy and equality violations; versus disgust for divinity and purity violations; Heerdink, Koning, van Doorn, & van Kleef, 2019; Shweder, Much, Mahapatra, & Park, 1997; see also, Sunar et al., 2021). However, different people experience different emotions toward moral violations with identical content (for a review, see Cameron, Lindquist, & Gray, 2015). For example, some individuals report greater anger toward moral violations, whereas others report greater disgust toward the same violation (Tybur, Molho, Cakmak, Cruz, Singh, & Zwicker, 2020b). Recent work has aimed to better understand these differences.

One category of proposals suggests that reports of anger toward moral violations are functionally equivalent to reports of disgust. This perspective seems consistent with findings suggesting that verbal self-reports of anger and disgust are nearly perfectly correlated (e.g., Russell & Giner-Sorolla, 2011). Some researchers have thus argued that expressions of (moral) disgust are actually expressions of anger or that people are confused about their emotional state (e.g., Alvarado, 1998; Nabi, 2002). Hence, according to this perspective, disgust toward moral violations is roughly (or entirely) equivalent to anger rather than a moral emotion with distinct antecedents and consequences.

Other work suggests that differences between anger and disgust are meaningful. Studies assessing anger and disgust via endorsements of canonical facial expressions rather than linguistic labels have revealed that disgust, more than anger, is activated in response to information about (bad) moral character (e.g., Giner-Sorolla & Chapman, 2017); that

☆ This paper has been recommended for acceptance by Pranjal Mehta.

* Corresponding author at: Room MF-C580, Van der Boechorststraat 7, 1081 BT Amsterdam, the Netherlands.
E-mail address: l.fan@vu.nl (L. Fan).

https://doi.org/10.1016/j.jesp.2024.104597
Received 10 January 2023; Received in revised form 10 January 2024; Accepted 12 January 2024

anger communicates more self-interest, whereas disgust communicates a more principled, moral motivation (Kupfer & Giner-Sorolla, 2017); and that disgust is activated more when moral violations target others than when they target the self or relatives, whereas anger is activated more when moral violations target the self or relatives than when they target others (Lopez, Moorman, Schneider, Baker, & Holbrook, 2021; Molho, Tybur, Guler, Balliet, & Hofmann, 2017). Another model suggests that, although expressions of disgust sometimes reflect anger (rather than a functionally distinct emotional response), moral disgust indeed has functions distinct from anger (Van der Eijk & Columbus, 2023).

1.1. Anger versus disgust and different types of aggression

In addition to the differences described above, anger and disgust also have distinct relations with motivations to punish, which are characterized by different types of aggression (e.g., Archer & Coyne, 2005; Molho, Twardawski, & Fan, 2022). One type – direct aggression – involves verbally or physically aggressing against a transgressor (e.g., Barker, Tremblay, Nagin, Vitaro, & Lacourse, 2006; Jambon & Smetana, 2018; Mathieson & Crick, 2010). Another type – indirect aggression – involves damaging a transgressor's reputation and recruiting others to punish by transmitting negative information about the transgressor to others (e.g., gossip; Feinberg, Cheng, & Willer, 2012; Wu, Balliet, & Van Lange, 2016). Findings suggest that anger is more strongly associated with motives to directly aggress, whereas disgust is more strongly associated with motives to indirectly aggress (Fan et al., 2023; Lopez et al., 2021; Molho et al., 2017; Molho, Tybur, Van Lange, & Balliet, 2020; Ocampo et al., 2022; Van der Eijk & Columbus, 2023).

These empirical findings align well with existing accounts of the functions of anger and disgust. Anger putatively motivates interventions that change another person's behavior in ways that place more weight on the angry person's interests (Fischer & Roseman, 2007; Sell et al., 2017; Sznycer, Sell, & Dumont, 2022; Sznycer, Sell, & Lieberman, 2021). Such behavioral changes are better executed via direct aggression, since the reason for such aggression, and its associated costs, is clear to the recipient. Various approaches – including diary studies (e.g., Molho et al., 2020), experimental studies (e.g., Molho et al., 2017; Wyckoff, 2016), and personality studies (e.g., Veenstra, Bushman, & Koole, 2018) – are consistent with this perspective.

Disgust toward moral violations does not appear to motivate the same physical avoidance as disgust toward pathogen cues (Kupfer & Giner-Sorolla, 2021). Yet it does not seem to motivate the same behaviors as anger toward moral violations either. Multiple studies indicate that compared to anger, disgust is less strongly related to direct aggression motives toward moral violators, but equally (or more) strongly related to indirect aggression motives (e.g., Molho et al., 2017). These results are consistent with proposals that disgust toward moral violations motivates ostracism, social distancing, or the recruitment of others for collective condemnation (Curtis & Biran, 2001; Hutcherson & Gross, 2011; Tybur, Lieberman, & Griskevicius, 2009; Tybur, Lieberman, Kurzban, & DeScioli, 2013). In contrast with direct aggression, which is more strongly associated with anger, indirect aggression is less effective at modifying the targets' behavior in a manner advantageous to the aggressor, since the reason for the aggression and the punisher's identity are less clear to the target. At the same time, the costs of this type of aggression are lower, since the aggressor does not risk (immediate) counter-aggression from the target (Archer & Coyne, 2005).

At their core, the studies described above were based on the idea that identical moral violations impose different costs on different individuals. When a moral violation imposes more costs on an individual, that individual will be more likely to experience anger, which is in turn associated with a more costly – but more immediately effective – direct type of aggression. When a moral violation imposes lower costs on an individual, that individual will be more likely to experience (moral) disgust, which is in turn associated with a less costly – but also less immediately effective – indirect type of aggression. These considerations lead to a straightforward hypothesis: a moral violation targeting oneself should elicit more anger than the same moral violation targeting another person, and a moral violation targeting another person should elicit more disgust than the same moral violation targeting oneself. Multiple studies have used just this approach (e.g., Molho et al., 2017; Tybur, Lieberman, Fan, Kupfer, & de Vries, 2020a). Other studies have suggested that moral violations targeting relatives are more costly than those targeting others, and hence disgust and anger should also vary differently depending on the identity of the "other" being targeted (Lopez et al., 2021). To date, though, research has only tested the logic above by comparing responses to moral violations targeting individuals that belong to specific folk categories, such as those of 'family' or 'friends.' Previous studies further assumed that individuals belonging to each category differ in their value to an observer—and hence track an observer's willingness to engage in costly emotional and aggressive action—though they did not directly test this idea. The arguments described above make a stronger prediction which we test here, which is that anger and disgust should relate differently to interpersonal value even within folk relationship categories.

1.2. Assessing interpersonal value and emotion

Interpersonal value varies across individuals even within similar folk categories of relationships (e.g., Tooby & Cosmides, 1996; Tybur et al., 2020b). Put simply: some people value their closest friend more so than other people do; some people value work acquaintances more than other people do; and some people value the person they like the least more than other people do. Existing research in this area has only manipulated folk categories that roughly vary in average interpersonal value. This practice leaves open the possibility that moral violations targeting more valued friends and acquaintances elicit the same emotional reactions as those targeting less valued friends and acquaintances – a possibility that would undermine current interpretations of previous findings. The studies presented here thus focused on testing whether the interpersonal value of a moral violation target relates differently to observers' anger versus disgust toward that violation. To achieve this, we examined responses to moral violations both across folk relationship categories (e.g., friend versus disliked person) and within such categories. In this way, we provide the first direct test of the idea that emotional reactions to moral violations track interpersonal value even within relationship types.

Further, previous work in this area has been limited by approaches to measuring anger and disgust, which carry substantial drawbacks although they are common in the emotion literature. Specifically, the widely-used method of the single-item verbal label rating task has been criticized for its limited reliability and validity (Weidman, Steckler, & Tracy, 2017). While the studies assessing differences between disgust and anger toward moral violations have largely assessed disgust via endorsement of posed facial expressions (cf. Piazza, Landy, Chakroff, Young, & Wasserman, 2018), they also have limitations. We mention three here. First, they have used single-item measures of endorsements of arrays of posed facial expressions, which have unclear (and presumably low) reliability. Second, posed facial expressions are sometimes interpreted by participants differently than intended by researchers (e.g., Piazza & Landy, 2020; Widen & Russell, 2008). Third, emotion expressions, as well as their decoding, vary across individuals (for a review, see Hildebrandt, Olderbak, & Wilhelm, 2015). Single-item endorsements of arrays force participants to decode each stimulus and summarize the common component of all presented expressions. The complexity of this task might further compromise the reliability and validity of the assessment. Study 2 addresses these limitations by using multi-modal assessments of anger and disgust.

1.3. Overview of the present studies

The current paper includes two studies which aim to test the relations between anger and disgust and aggressive behaviors toward moral violators. It extends previous work by assessing the interpersonal value of the target of the moral violation and by using multi-modal assessments of anger and disgust.

Study 1 employed a between-subject design in which participants read about a moral violation that targeted either a close friend, an acquaintance, or someone they know but dislike. This manipulation was intended to increase the range of interpersonal value among moral violation targets. Participants then completed a financial tradeoff task intended to estimate the interpersonal value of the target and reported their aggression motives (both direct and indirect) and emotions (anger and disgust) toward the moral violator. Study 2 improved upon this approach in three ways. First, rather than assessing anger and disgust based only on agreement with a single array of faces, we used endorsement of multiple individual facial and vocal expressions. Second, we manipulated moral violation targets' interpersonal value within-participants across multiple experimental sessions. Third, to avoid any residual effects of folk-label categories (e.g., "friend"), we facilitated variation in target interpersonal value by having participants think of a target at a specific social distance from them.

Both studies used a stimulus sampling approach in which participants were randomly assigned to read one of 12 moral violation scenarios. We aimed to test (1) whether a target's interpersonal value relates more strongly to anger and direct aggression, but less so to disgust and indirect aggression, and (2) whether anger relates more strongly to direct aggression than indirect aggression, whereas disgust relates more strongly to indirect aggression than direct aggression.

Participants in Study 1 were recruited via Amazon MTurk, while participants in Study 2 were students recruited from Vrije Universiteit Amsterdam. Preregistrations, materials, data, and analysis scripts for both studies are available on the Open Science Framework (https://osf.io/36zxr/).

2. Study 1

2.1. Method

2.1.1. Participants

Data were collected via Amazon MTurk in June 2019. According to an a-priori power analysis using R 4.0.3 (R Core Team, 2020) with the package SimR (Green & MacLeod, 2016), a sample of 834 participants affords 80% power to detect an interaction between interpersonal value (represented by Welfare Tradeoff Ratio, WTR) and type of aggression motive (direct versus indirect) ($d = 0.4$). We anticipated excluding approximately 8% of responses. Hence, we targeted a sample size of 908 participants.

As pre-registered, we excluded 24 responses due to participants providing nonsensical free-text responses (as identified by two coders) and 39 participants with more than two switch points for any of the three WTR anchors (see Welfare Tradeoff Task section, for more details). A total of 847 valid responses (60% male, $M_{age} = 36.88$, $SD = 11.38$) were included in the analysis.

2.1.2. Procedures

Participants first completed measures of HEXACO Honesty-Humility and Agreeableness. They then were randomly assigned to picture either a close friend, an acquaintance, or a disliked person and write down a brief description of this person (hereafter referred to as "the target") as well as this person's initials or nickname. Afterward, participants completed the Welfare Tradeoff Task (WTT) toward the target.

Participants were then randomly assigned to read one of the 12 scenarios describing a moral violation against the target. The target's initials or nickname were integrated in the presentation of the moral violation. After that, participants reported the degree to which their emotional reaction to the scenario corresponded with one array of faces expressing anger and another array of faces expressing disgust, how morally wrong they found the violation, and the degree to which they felt like engaging in several aggressive acts toward the moral violator, some of which were directly aggressive, and others of which were indirectly aggressive. We also collected participants' demographic information. Participants were paid $2 for their participation.

2.1.3. Instruments and materials

Honesty-Humility and Agreeableness. We used the Honesty-Humility and Agreeableness scales from the HEXACO-60 (Ashton & Lee, 2009). Previous studies showed that some pro-social personality traits are positively correlated with WTR and negatively correlated with aggressive behaviors (Dinic & Wertag, 2018; Kirkpatrick, Delton, Robertson, & de Wit, 2015; Knight, Dahlen, Bullock-Yowell, & Madson, 2018). Hence, we controlled for these two personality variables.

Welfare Tradeoff Task. We used the WTT to measure interpersonal value (Delton, 2010; Pedersen, 2015; Smith, Pedersen, Forster, McCullough, & Lieberman, 2017). The task instructs participants to indicate their preference between receiving a specific amount of money for themselves versus having the target receive a specific amount of money. WTR is estimated as the value at which participants switch from preferring benefits to themselves to preferring benefits to the target.

We used a 30-item, three-anchor version of the WTT, with anchor amounts of $19, $46, and $75 (see online supplement for further details). Each anchor included 10 values, which ranged from −0.35 times the anchor point value (e.g., $–26 for the $75 anchor point) to 1.45 times the anchor point value (e.g., $109 for the $75 anchor point). The switch point was the average of the last amount the participant selected for him/herself, and the first amount forgone to give resources to the other person. Consider a participant who, for the $75 anchor, always chooses to receive the money when the potential gain is greater than the anchor (i.e., for values of $79, $94, and $109). Conversely, he/she opts to give the money to the other person when the potential gain is smaller than the anchor (i.e., for $64, $49, $34, $19, $4, and the two situations of losing $11 and $26). In this example, WTR is the midpoint of the two amounts where this shift happens (in this case, the average of $79 and $64, which is $73) divided by the anchor value (here, $75), so 0.95. Participants who always chose the other-benefiting option received a value of 1.55 for that anchor; those who never chose the other-benefiting option received a value of −0.45. If there were two switch points within an anchor, these values were averaged; participants with more than two switch points for any anchor were excluded. WTR was estimated as the average of the switch points for each anchor.

Moral violation scenarios. Inspired by Study 4 of Molho et al. (2017), we generated a set of 12 scenarios, each of which described a different moral violation affecting the target (e.g., property damage, such as damaging a coat with cigarette ash; theft, such as stealing a wallet from a coat; or assaulting, such as physically slapping someone who acts inconsiderately). The target's initials or nickname were included as the target in the scenario. Scenarios were designed to be similar in length; they ranged from 111 words to 142 words.

Emotion endorsement and moral wrongness evaluation. As in Molho et al. (2017), we assessed anger and disgust by presenting one array of six posed anger faces and one array of six posed disgust faces retrieved from the Radboud Faces Database (RaFD, Langner et al., 2010) and asking participants to rate how well these two sets of facial expressions matched their feeling ("These faces match how I felt when I read the scenario") on a seven-point scale (1 = strongly disagree, 7 = strongly agree). We also asked participants "how morally wrong do you think the behavior of the person in this scenario was?" (0 = not morally wrong at all, 100 = extremely morally wrong).

Aggression. We asked participants to rate the degree to which they agreed with statements describing their potential aggressive responses toward the perpetrator on a seven-point scale (1 = strongly disagree, 7

= strongly agree). The items were retrieved from Molho et al. (2017) and revised to fit the scenarios. According to an exploratory factor analysis, one direct aggression item (loadings: 0.34, 0.59) and two indirect aggression items (loadings: 0.41, −0.38; 0.43, 0.24) were excluded due to their low or less-distinctive loadings (for details see SOM). Direct and indirect aggression scores were calculated by taking the mean of the remaining items (α = 0.90 for direct aggression, α = 0.91 for indirect aggression).

2.1.4. Analysis

Given that we used different target labels to facilitate variation in WTR, we first verified that WTRs indeed differed across these labels. To test the relationships between interpersonal value and aggression motives and interpersonal value and emotion endorsements, we respectively examined interactions between WTR and aggression type and WTR and emotion type. To test the patterns of relations between the two emotions and different aggression motives, we tested the interaction between anger and the two aggression types and the interaction between disgust and the two aggression types. We only pre-registered controlling for target label in all these models. Based on suggestions during the peer-review process, we also (1) report analyses showing differences in emotion and aggression across target label categories, (2) control for interactions between target label categories and types of aggression and emotion, and (3) examine how anger and disgust mediate relations between target features and aggression. Finally, per our preregistration, we also analyzed models controlling for personality variables. We report key results in the main text. Other analyses are included in the SOM.

For all models, we included scenario and participant as random intercepts. We initially fitted the models with the main effects underlying the relevant interaction, followed by a second model adding the interaction terms, a third model controlling for participant sex, target sex, participant age, and perceived moral wrongness, and a fourth model controlling for agreeableness and honesty-humility. All significant interactions were probed with lower-order simple-effect tests. In the main text, we report the results of the final models with all control variables. For results of each step, see SOM.

2.2. Results

2.2.1. Manipulation check and preliminary analyses

Participants reported the highest WTR when the target was a close friend (0.68, 95% CI [0.63, 0.72]), followed by when the target was an acquaintance (0.34, 95% CI [0.29, 0.38]), followed by when the target was a disliked person (−0.06, 95% CI [−0.11, −0.02], $F(2,884) = 306.00$, $\eta_p^2 = 0.41$, $p < .001$). Variation in interpersonal value was substantial within each of these three target labels ($SD_{friend} = 0.39$, $SD_{acquaintance} = 0.39$, $SD_{disliked} = 0.29$). Bivariate correlations and Cronbach's alphas (for multi-item measures) are reported in Table 1.

2.2.2. Interpersonal value and emotion endorsements

We observed a significant interaction between WTR and emotion type ($\beta = 0.09$, 95% CI [0.06, 0.12], $t(845) = 5.56$, $p < .001$). Target WTR related more strongly to anger ($\beta = 0.35$, 95% CI [0.29, 0.41]) than disgust ($\beta = 0.17$, 95% CI [0.11, 0.23], see Fig. 1a left). Differences between anger and disgust also varied across relationship labels ($F(2, 844) = 20.02$, $\eta_p^2 = 0.05$, $p < .001$). The simple effect of relationship label was stronger for anger ($F(2, 1439) = 100.07$, $\eta_p^2 = 0.12$, $p < .001$; $M_{friend} = 5.60$, 95% CI [5.34, 5.85]; $M_{acquaintance} = 4.93$, 95% CI [4.67, 5.18]; $M_{disliked} = 3.62$, 95%CI [3.36, 3.87]) than for disgust ($F(2, 1439) = 27.86$, $\eta_p^2 = 0.04$, $p < .001$; $M_{friend} = 4.65$, 95% CI [4.40, 4.91]; $M_{acquaintance} = 4.14$, 95% CI [3.89, 4.40]; $M_{disliked} = 3.59$, 95%CI [3.33, 3.84]).

We next tested a model including both WTR and two orthogonally coded variables representing the three relationship labels. The interaction between WTR with emotion type remained statistically significant after controlling for relation labels ($\beta = 0.09$, 95% CI [0.06, 0.12], $t(845) = 5.56$, $p < .001$), and also after controlling for the interaction between label condition and emotion type ($\beta = 0.05$, 95% CI [0.01, 0.09], $t(843) = 2.21$, $p = .03$).

2.2.3. Interpersonal value and aggression motives

We observed an interaction between WTR and type of aggression motive ($\beta = 0.05$, 95% CI [0.02, 0.08], $t(845) = 3.16$, $p = .002$, see Fig. 1a right). The relation between WTR and direct aggression ($\beta = 0.25$, 95% CI [0.19, 0.31]) was stronger than that between WTR and indirect aggression ($\beta = 0.16$, 95% CI [0.10, 0.22]). A similar pattern emerged for relationship labels ($F(2, 844) = 7.70$, $\eta^2_{partial} = 0.02$, $p < .001$; for direct aggression: $M_{friend} = 3.73$, 95% CI [3.47, 4.00], $M_{acquaintance} = 3.02$, 95% CI [2.75, 3.28], $M_{disliked} = 2.28$, 95%CI [2.01, 2.54]; for indirect aggression: $M_{friend} = 3.73$, 95% CI [3.47, 4.00], $M_{acquaintance} = 3.48$, 95% CI [3.21, 3.75], $M_{disliked} = 2.64$, 95%CI [2.38, 2.91]). The interaction between WTR and aggression type remained after controlling for relation labels ($\beta = 0.05$, 95% CI [0.02, 0.08], $t(845) = 3.16$, $p = .002$). When further including the interaction between relationship labels and aggression type, the interaction was non-significant ($\beta = 0.03$, 95% CI [−0.00, 0.07], $t(843) = 1.76$, $p = .08$).

2.2.4. Emotion endorsements and aggression motives

Finally, we tested the effects of the anger and disgust endorsements on aggression motives. Results revealed significant interactions between anger and aggression type ($\beta = -0.06$, 95% CI [−0.10, −0.03], $t(844) = -3.64$, $p < .001$) and between disgust and aggression type ($\beta = 0.07$, 95% CI [0.04, 0.10], $t(844) = 4.06$, $p < .001$). Anger related more strongly to direct aggression ($\beta = 0.42$, 95% CI [0.36, 0.49]) than indirect aggression ($\beta = 0.28$, 95% CI [0.21, 0.35]), while disgust related more strongly to indirect aggression ($\beta = 0.16$, 95% CI [0.09, 0.22]) than direct aggression ($\beta = 0.03$, 95% CI [−0.03, 0.09], see Fig. 2a).

Table 1
Bivariate correlations between Study 1 variables.

Variables	1	2	3	4	5	6	7	8
1 Direct aggression	0.91							
2 Indirect aggression	***0.62***	0.86						
3 Anger	***0.47***	***0.43***						
4 Disgust	***0.27***	***0.35***	***0.51***					
5 WTR	***0.27***	***0.20***	***0.38***	***0.20***				
6 Moral Wrongness	***0.29***	***0.28***	***0.42***	***0.27***	***0.12***			
7 Agreeableness	***−0.15***	***−0.23***	−0.03	−0.02	−0.01	0.01	0.84	
8 Honesty-Humility	***−0.19***	***−0.25***	0.04	−0.06	*0.09*	*0.07*	***0.31***	0.81
9 Participant Sex	***0.20***	**0.10**	−0.03	−0.05	0.07	−0.02	0.05	***−0.16***
10 Target Sex	**−0.11**	−0.04	−0.07	*−0.07*	−0.06	***−0.12***	−0.00	−0.01

Note: ***Bold and italics*** = $p < .001$, **bold** = $p < .01$, *italics* = $p < .05$. Cronbach's alphas of multi-item measurements are on the diagonal. For target and participant sex, 1 = male and 0 = female. Correlations equaling to zero indicate a correlation smaller than 0.005.

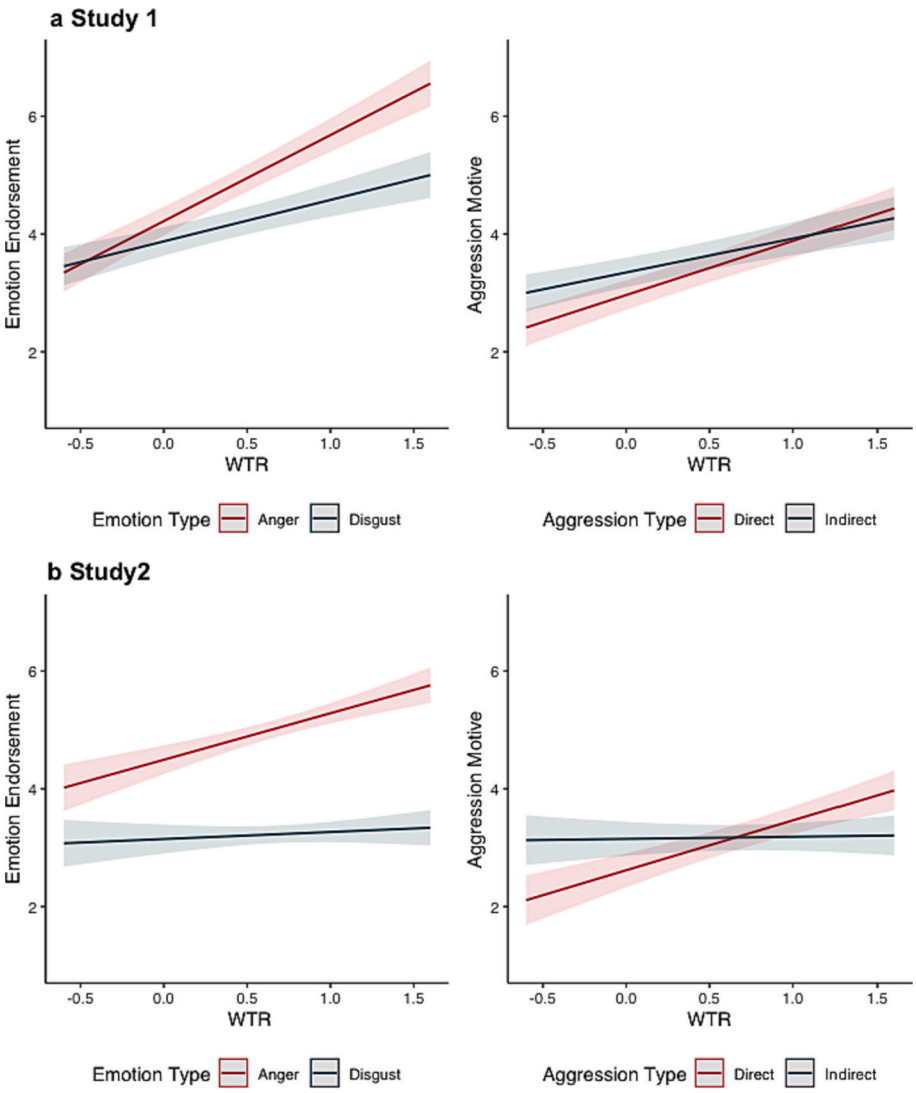

Fig. 1. Marginal effects of WTR predicting specific moral emotion endorsements and aggression motives in Studies 1 (a) and 2 (b).
Note: The panels on the left side indicate the simple slopes of WTR on emotion endorsements, and the panels on the right side indicate those for aggression motives. The shaded areas indicate the 95% CI of the slopes. The models include random intercepts for moral violation scenarios and participants and fixed effects of participant sex, target sex, participant age, perceived moral wrongness, Agreeableness, and Honesty-Humility.

2.2.5. Emotion endorsements as mediators

Using the lavaan package (Rosseel, 2012), we fitted models with either WTR (the first model) or relationship labels (the second model) as exogenous variables, anger and disgust endorsements as mediators, and direct and indirect aggression as dependent variables (see Fig. S4 in SOM for the detailed results).

The first model revealed that anger partially mediated relations between WTR and both direct aggression ($\beta = 0.15$, 95% CI [0.11, 0.19], $p < .001$) and indirect aggression ($\beta = 0.12$, 95% CI [0.09, 0.16], $p < .001$), whereas disgust partially mediated the relation between WTR and indirect aggression only ($\beta = 0.04$, 95% CI [0.02, 0.06], $p < .001$). In the second model, we set the acquaintance condition as the reference. Similar to the first model, anger partially mediated the relation between relationship labels and both direct aggression ($\beta_{Friend} = 0.15$, 95% CI [0.11, 0.19], $p < .001$; $\beta_{Disliked} = -0.18$, 95% CI [−0.23, −0.13], $p < .001$) and indirect aggression ($\beta_{Friend} = 0.11$, 95% CI [0.07, 0.15], $p < .001$; $\beta_{Disliked} = -0.13$, 95% CI [−0.18, −0.09], $p < .001$), whereas disgust only mediated the relation between relationship labels and indirect aggression ($\beta_{Friend} = 0.05$, 95% CI [0.02, 0.07], $p < .001$; $\beta_{Disliked} = -0.05$, 95% CI [−0.07, −0.02], $p < .001$).

2.3. Discussion

In Study 1, we tested whether the interpersonal value of a target of a moral violation related differently to anger versus disgust and direct versus indirect aggression toward a perpetrator. As hypothesized, interpersonal value related more strongly to endorsements of anger than disgust and more strongly to direct aggression than indirect aggression. Further, anger related more strongly to direct aggression than indirect aggression, whereas disgust related more strongly to indirect aggression than direct aggression.

This study replicates and extends findings from previous studies on the functions of anger and disgust toward moral violations (e.g., Lopez et al., 2021; Molho et al., 2017) by assessing interpersonal value rather than only comparing reactions to moral violations targeting the self and others or kin and acquaintances. Nevertheless, Study 1 was limited by the same single-item assessment of emotion used in earlier research. Further, in Study 1, we aimed to capture a wide range of interpersonal value by prompting participants to think about a close friend, an acquaintance, or someone they really disliked. Folk categories such as "friend" might produce demand characteristics (e.g., based on expectations of behavior toward others conceived of as friends) which might

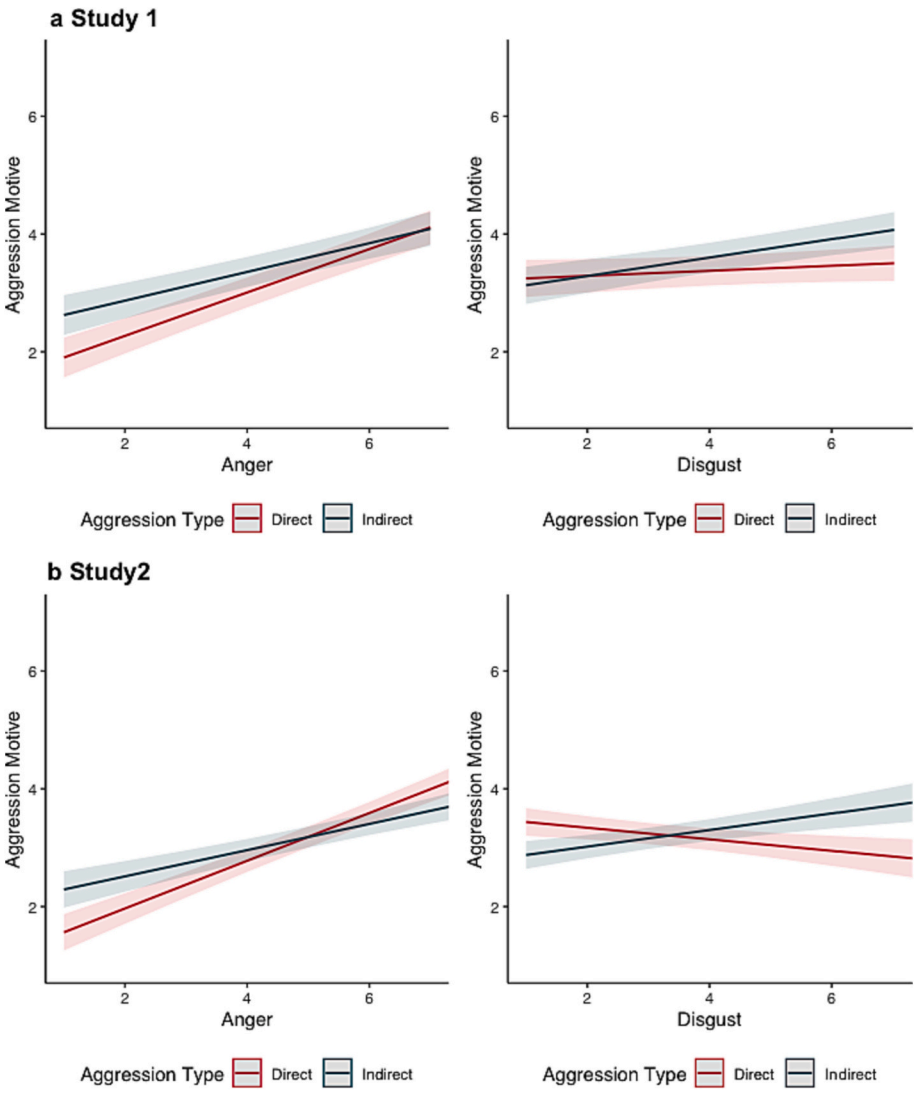

Fig. 2. Marginal effects of moral emotions predicting aggression motives in Studies 1 (a) and 2 (b).
Note: The shaded areas indicate the 95% CI of the slopes. The models include random intercepts for moral violation scenarios and participants and fixed effects of participant sex, target sex, participant age, perceived moral wrongness, Agreeableness, and Honesty-Humility.

have influenced emotional and aggressive responses to moral violations independent of interpersonal value.

Study 2 addressed these limitations by: (1) using a multi-modal and multi-item assessment of anger and disgust; and (2) using a manipulation of interpersonal value of targets that does not rely on folk relationship categories. Additionally, we manipulated the identity of the moral violation target within-participants across multiple experimental sessions rather than using the between-participant approaches of other studies in this literature (e.g., Lopez et al., 2021; Molho et al., 2017).

3. Study 2

3.1. Method

3.1.1. Participants

Students at Vrije Universiteit Amsterdam were recruited between fall 2021 and spring 2022. According to an a-priori power analysis based on effect sizes from Study 1, a sample of 220 participants afforded 85% power to detect an interaction between WTR and the aggression motive type ($d = 0.20$, for scripts see online supplements on OSF). Given opportunities for online data collection afforded by the COVID-19 pandemic, we collected as many participants as possible ($N = 365$) before the end of the university term. We excluded five participants who did not identify as either men or women given our pre-registered plan to control for participant sex. The final sample included 360 valid responses (23% male, $M_{age} = 20.49$, $SD = 2.39$). Participants were compensated with course credits.

We offered the surveys in English and Dutch. Native Dutch-speaking participants were directed to the Dutch version of the survey, and other participants completed the English version. For the Dutch version, all the materials except the HEXACO-60 (for which there is a published Dutch version, De Vries, Lee, & Ashton, 2008) were translated from English to Dutch by a bilingual native Dutch speaker, back-translated into English by a bilingual native Dutch speaker, and checked by a native English speaker for consistency with the original materials.

3.1.2. Procedures

As in Study 1, participants first provided informed consent. In the consent process, they were told that the study included four sessions, one administered per week.

In each session, participants imagined a person they know (see below for more details) and then completed the WTT involving the target. Participants then read one of the vignettes used in Study 1 in which the target was described as the victim of a moral violation. Participants then

completed the emotion endorsement, moral wrongness evaluation, and aggression items. We also collected demographic information (e.g., age, sex), and some other variables for exploratory purposes (see OSF supplement for an exhaustive list of measures).

Target identities and scenarios were assigned in a pseudo-random order for each participant between sessions.[1] As in Study 1, target initials or nicknames were integrated in the text of WTT and moral violation vignettes.

3.1.3. Instruments and materials

Most instruments were identical to those used in Study 1. We describe two exceptions: the materials used to facilitate variation in target WTR and the materials used to assess anger and disgust.

Target manipulation. To facilitate variation in target WTR, we used a method based on Jones and Rachlin (2006). Participants were asked to imagine a list of the 100 people closest to them ranging from their dearest friend or relative at #1 to a mere acquaintance at #100. They were randomly assigned to think of a person at #1, #10, #20, or #40. As in Study 1, they were asked to write the person's initials or nickname, to provide some basic information about the person, and to give a brief description of the targeted person in a free response box. Afterward, they completed the WTT toward the target.

Emotion. We aimed to improve upon our assessment of emotion in two ways. First, in addition to assessing agreement with posed facial expressions, we also assessed agreement with non-verbal vocal expressions. Second, instead of using a single-item measure of agreement with an array of stimuli, we asked participants to rate how well individual facial and vocal stimuli matched their feeling on a seven-point scale (1 = strongly disagree, 7 = strongly agree). Facial expressions were retrieved from RaFD (Langner et al., 2010) and vocal expressions were retrieved from Sauter, Eisner, Calder, and Scott (2010) and Yoshie and Sauter (2019) (also see Fan & Tybur, 2021, and Fan et al., 2023). We selected 24 items in total: six anger faces, six disgust faces, six anger vocalizations, and six disgust vocalizations. The average of agreement with the six anger faces ($\alpha = 0.91$) correlated strongly with the average of agreement with the six anger vocalizations ($\alpha = 0.91$), $r = 0.70$, $p < .001$, and the average of agreement with the six disgust faces ($\alpha = 0.92$) correlated strongly with the average of the six disgust vocalizations ($\alpha = 0.92$), $r = 0.59$, $p < .001$. Correlations within modalities but across emotions were much lower ($r = 0.33$, $p < .001$ for vocalizations; $r = 0.24$, $p < .001$ for faces). We therefore took the average of anger across modalities and the average of disgust across modalities. No conclusions changed when either vocal or face measures were used individually.

3.1.4. Analysis

We followed the same data exclusion and analysis plan used in Study 1.

3.2. Results

3.2.1. Manipulation check and preliminary analyses

We first checked whether WTR varied as a function of the target manipulation. Participants reported the highest WTR when the targets ranked #1 (0.81, $SD = 0.30$, 95%CI [0.77, 0.86]), followed by #10 (0.73, $SD = 0.34$, 95%CI [0.68, 0.77]), #20 (0.63, $SD = 0.32$, 95%CI [0.58, 0.67]), and #40 (0.55, $SD = 0.35$ 95%CI [0.50, 0.59], $F(3, 847) = 26.60$, $\eta_p^2 = 0.09$, $p < .001$). Bivariate correlations are reported in Table 2.

3.2.2. Interpersonal value and emotion endorsements

We again detected an interaction between target WTR and emotion type ($\beta = 0.07$, 95% CI [0.03, 0.10], $t(1349) = 3.71$, $p < .001$, see Fig. 1b left). Target WTR related more strongly to anger ($\beta = 0.16$, 95% CI [0.10, 0.21]) than disgust ($\beta = 0.02$, 95% CI [−0.03, 0.08]). We did not observe a similar interaction across rank labels ($F(3, 1357 = 2.50$, $\eta_p^2 = 0.01$, $p = .06$). The interaction between WTR and emotion type remained when controlling for rank labels ($\beta = 0.07$, 95% CI [0.03, 0.10], $t(1349) = 3.71$, $p < .001$) and the interaction between rank labels and emotion type ($\beta = 0.06$, 95% CI [0.02, 0.09], $t(1343) = 3.08$, $p = .002$).

3.2.3. Interpersonal value and aggression motives

The effect of WTR on aggression motives again differed according to aggression types ($\beta = 0.09$, 95% CI [0.05, 0.13], $t(845) = 4.52$, $p < .001$, see Fig. 1b right). WTR related to motivations to directly aggress ($\beta = 0.19$, 95% CI [0.13, 0.26]), but not to motivations to indirectly aggress ($\beta = 0.01$, 95% CI [−0.05, 0.07]). In contrast, the interaction between rank labels and aggression type was not significant ($F(3, 1359) = 1.41$, $\eta^2_{partial} = 0.003$, $p = .24$). The interaction between WTR and aggression type remained when controlling for the rank labels ($\beta = 0.09$, 95% CI [0.05, 0.13], $t(1347) = 4.54$, $p < .001$) and the interaction between rank labels and emotion type ($\beta = 0.09$, 95% CI [0.05, 0.13], $t(1344) = 4.46$, $p < .001$).

3.2.4. Emotion endorsements and aggression motives

Regressing anger and disgust on aggression revealed that disgust and anger differentially related to direct and indirect aggression ($\beta_{anger} = 0.09$, 95% CI [0.05, 0.13], $t(1361) = 4.61$, $p < .001$; $\beta_{disgust} = -0.11$, 95% CI [−0.15, −0.07], $t(1361) = -5.53$, $p < .001$). Anger related more strongly to direct aggression ($\beta = 0.41$, 95% CI [0.35, 0.48]) than indirect aggression ($\beta = 0.23$, 95% CI [0.16, 0.29]), whereas disgust related positively to indirect aggression ($\beta = 0.13$, 95% CI [0.07, 0.20]) but negatively to direct aggression ($\beta = -0.09$, 95% CI [−0.15, −0.03], see Fig. 2b).

3.2.5. Emotion endorsements as mediators

The first model, which assessed how emotion mediates relations between WTR and aggression, indicated that anger partially mediated the relation between WTR and both direct ($\beta = 0.07$, 95% CI [0.04, 0.10], $p < .001$) and indirect aggression ($\beta = 0.05$, 95% CI [0.03, 0.07], $p < .001$), whereas disgust did not mediate the relation between WTR and either aggression type. The second model, which assessed how emotion mediated relations between target labels and aggression, revealed that anger partially mediated only the contrast between #1 and #40 for both direct ($\beta = -0.05$, 95% CI [−0.08, −0.01], $p < .01$) and indirect aggression ($\beta = -0.03$, 95% CI [−0.05, −0.01], $p < .05$, see Fig. S4 in SOM for more detailed results).

3.3. Discussion

In Study 2, we again tested the hypothesis that anger versus disgust and direct versus indirect aggression toward a moral violator vary as a function of the interpersonal value of the person targeted by that violation. We did so with a better assessment of anger and disgust than that used in Study 1 and those used in similar studies in this area (e.g., Molho et al., 2017; Tybur et al., 2020b). Findings again indicated that people report greater anger toward moral violations targeting more interpersonally valued others, and that the strength of this relation is greater than the strength of the relation between disgust and the value placed on a target. Indeed, in Study 2 – unlike in Study 1 – the relation between disgust and target interpersonal value did not differ from zero.

[1] We used participants' birth month as an indicator for pseudo-randomization in the manipulation. Participants were divided into four groups based on this indicator (e.g., Group 1 consisted of those born in January, April, or August) and were then assigned a specific sequence of ranking manipulation across sessions (e.g., in Group 1, participants experienced rankings of #40, #20, #1, and #60 sequentially). This method was similarly employed for scenario assignments. However, we used a different birth month grouping strategy. For each group in each session, three scenarios were potential candidates, with only one being randomly presented to each participant.

Table 2
Bivariate correlations between Study 2 variables.

Variables	1	2	3	4	5	6	7	8
1 Direct aggression	0.85							
2 Indirect aggression	***0.35***	0.87						
3 Anger	***0.40***	***0.30***	0.93					
4 Disgust	0.01	***0.19***	***0.29***	0.93				
5 WTR	***0.13***	−0.05	***0.17***	0.02	–			
6 Moral Wrongness	***0.18***	***0.16***	***0.25***	−0.00	***0.17***	–		
7 Agreeableness	***−0.16***	***−0.12***	0.01	−0.00	*0.09*	0.04	0.75	
8 Honesty-Humility	***−0.14***	***−0.18***	0.04	−0.02	***0.21***	*0.08*	***0.30***	0.66
9 Participant Sex	***0.18***	−0.03	*0.08*	−0.01	−0.03	−0.06	*0.07*	***−0.22***
10 Target Sex	0.02	−0.05	−0.00	−0.01	−0.02	*−0.07*	0.02	−0.04

Note: ***Bold and italics*** = $p < .001$, **bold** = $p < .01$, *italics* = $p < .05$. Cronbach's alphas of multi-item measurements are on the diagonal. For target and participant sex, 1 = male and 0 = female. Correlations equaling to zero indicate a correlation smaller than 0.005.

Like in Study 1, anger related more strongly to direct than indirect aggression, whereas disgust related more strongly to indirect than direct aggression.

4. General discussion

Why do people who witness the same moral violations have different emotional and behavioral responses toward that violation? Two studies suggest that part of this variation reflects the value that an individual places on the target of the moral violation. Specifically, the value placed on a target relates more strongly to anger than to disgust, and more strongly to direct aggression than indirect aggression. We also observed consistent relations between emotion and aggression: anger endorsements related positively to both direct and indirect aggressive motives (and more so to direct aggression), whereas disgust related positively only to indirect aggressive motives. These results delineate the impact of interpersonal relationships on third-party punishment of moral violations and shed light on the functions of anger and moral disgust, including their associations with different forms of aggression.

4.1. Revisiting moral disgust

Aligned with previous studies (e.g., Lopez et al., 2021; Molho et al., 2017), we repeatedly observed that disgust and anger toward moral violations differentially related to distinct aggressive motives (see Table 3). These results suggest that disgust toward moral violations might have functions unique from those associated with anger – motivating indirect punishment tactics, such as gossip and social exclusion, but not motivating the direct aggression associated with anger (e.g., Molho et al., 2020; Molho & Wu, 2021). Results also support the proposition that situational variation in interpersonal relationships is associated with the experience of distinct moral emotions. To illustrate, anger increased along with the higher interpersonal value of the target, whereas disgust was less sensitive to interpersonal value and tended to be stronger when the interpersonal value of the target was lower. This finding aligns with previous work suggesting that people experience higher anger when they are personally victimized by offenses than when they are not, whereas they experience more disgust when they are not personally victimized than when they are (e.g., Molho et al., 2017). In sum, findings lend support to the suggestion of a cost-benefit mapping function of these two emotions (e.g., Molho et al., 2017; Sznycer et al., 2022; Tybur et al., 2013). Anger might function to terminate and deter transgressions, whereas disgust might instead function to coordinate condemnation, collective punishment, and social exclusion.

Table 3
Summary of key interaction results across Studies 1 and 2.

	Study 1		Study 2	
	Effect size (β)	95% CI	Effect size (β)	95% CI
WTR by emotions	0.09	[0.06, 0.12]	0.07	[0.03, 0.10]
- Anger	0.35	[0.29, 0.41]	0.16	[0.10, 0.21]
- Disgust	0.17	[0.11, 0.23]	*0.02*	[−0.03, 0.08]
WTR by aggression motives	0.05	[0.02, 0.08]	0.09	[0.05, 0.13]
- Direct aggression	0.25	[0.19, 0.31]	0.19	[0.13, 0.26]
- Indirect aggression	0.16	[0.10, 0.22]	*0.01*	[−0.06, 0.07]
Anger by aggression motives	0.07	[0.04, 0.10]	0.09	[0.05, 0.13]
- Direct aggression	0.42	[0.36, 0.50]	0.41	[0.35, 0.48]
- Indirect aggression	0.28	[0.22, 0.35]	0.23	[0.16, 0.29]
Disgust by aggression motives	−0.06	[−0.10, −0.03]	−0.11	[−0.15, −0.07]
- Direct aggression	*0.03*	[−0.03, 0.09]	−0.09	[−0.15, −0.03]
- Indirect aggression	0.16	[0.09, 0.22]	0.13	[0.07, 0.20]

Note: *Italics* = non-significant. All models were fit with random intercepts of participants and scenarios. For reported models involving WTR in the table, target label variables and their interactions with emotion and aggression type were not included.

4.2. Interpersonal value and third-party punishment

The distinct relations between interpersonal value and motivations to directly versus indirectly aggress suggest a possible tradeoff between the benefits and costs of aggression in third-party punishment contexts (see Table 3). Compared to indirect aggression, direct action might more effectively change a transgressor's behavior. Via physical or verbal aggression, punishers impose costs that the target is aware of, or they can signal their willingness to do so. These effects come at a cost: an increased likelihood of retaliation, a possible escalation of violence (Cushman, 2015; about violence escalation, see: DeWall, Anderson, & Bushman, 2011), and reputational costs associated with appearing violent (e.g., Eriksson, Andersson, & Strimling, 2016; Raihani & Bshary, 2015). Multiple features of indirect aggression render it less effective at altering a transgressor's behavior. First, the costs incurred by the transgressor are less immediate. Such delays make it harder for the transgressor to relate any reputational damage and social exclusion to the initial transgression. With such ambiguity, indirect aggression would not change behavior as effectively as direct aggression. However, these disadvantages could be outweighed by decreased costs to the punisher. Take gossip as an example. People spread gossip not to the target of that gossip, but to others. By doing so, gossip can achieve consensus and build alliances against the target. By using collective power, the costs of changing transgressors' behavior will be shared by group members. Meanwhile, the identity of each individual in the alliance is protected (or, at least, ambiguous), which leads to a lower retaliation possibility.

When the benefit of using direct aggression – in terms of inducing an immediate change in the transgressor's behavior – is sufficient to outweigh its costs, then direct aggression should be more likely. When the cost of direct aggression is higher than its benefits, indirect

aggression is more advantageous. The benefits of these approaches hinge partially on the target of a moral violation. When a moral violation targets a high-value other, the benefits of intervening are higher; when a moral violation targets a low-value other, the same benefits are lower.

Our focus on interpersonal value rather than only folk relationship categories has further implications for understanding the relational nature of third-party punishment. In both studies, interpersonal value related to both emotion and aggression independently of folk relationship categories (friend, acquaintance, disliked person) or ranking labels (#1, #10, #20, #40). Moral violations targeting more valued friends corresponded with different emotional and aggressive responses toward perpetrators than identical moral violations targeting less valued friends. Similarly, moral violations targeting more valued disliked persons elicited different patterns of emotional and aggressive responses than violations targeting more devalued disliked persons. This finding suggests that whether the target of a moral violation is considered to be a friend or a foe is not sufficient to understand reactions to that moral violation. That said, in Study 1, there were residual effects of relationship labels above and beyond interpersonal value. These residual effects were not observed in Study 2, when participants were asked to think of a target without the use of folk labels. These differences might have resulted from a mix of demand characteristics after having associated a target with a specific label ("friend"), a restriction in the range of interpersonal value in Study 2, and imprecision in assessments of interpersonal value via the welfare tradeoff tasks.

We also conducted mediation analyses to examine whether the observed effects of target interpersonal value on aggression can be partly attributed to emotional responses to moral violations. In both studies, anger partially mediated the relation between interpersonal value and both direct and indirect aggression, whereas disgust partially mediated only the relation between interpersonal value and indirect aggression, and only in Study 1. These findings are broadly consistent with patterns observed in previous studies (Lopez et al., 2021; Molho et al., 2017), although those studies focused on different types of moral violation targets (self versus other, or self versus sibling versus friend, respectively). In sum, findings suggest that indirect effects of the target of moral violations on aggression are more consistent via anger (as in Lopez et al., 2021, Studies 2 and 3). Further, they tentatively suggest that disgust only mediates the relation between moral violation target and indirect aggression (as in Molho et al., 2017, Study 4; though here, only in Study 1). As with these earlier studies that have conducted similar mediation analyses, inferences drawn from these analyses should remain tentative, especially because the mediators in the models (anger and disgust) were observed rather than experimentally manipulated (Fiedler, Harris, & Schott, 2018; Rohrer, Hünermund, Arslan, & Elson, 2022). Ultimately, the bivariate relations reported here offer the best information for theory development.

4.3. Assessments of anger and disgust

Much research in this area relies upon assessments of emotion with unclear validity (Weidman et al., 2017). In Study 1, as in a handful of previous studies examining relations between anger, disgust, and aggression (Lopez et al., 2021; Molho et al., 2017), we assessed emotion by asking participants the degree to which a single array of faces expressing anger matched their feelings and a single array of faces expressing disgust matched their feelings. Although, in this approach, endorsements of anger and disgust are more modestly correlated than when using verbal self-reports of anger and disgust (Gutierrez & Giner-Sorolla, 2007), the validity of this approach remains unclear. In Study 2, we instead assessed agreement with six individual faces and six individual vocal tokens for each of the two emotions. This approach revealed strong correlations within emotion but across modality and weak correlations across emotion but within modality. We also tested the improvement in reliability. Opting for a single-item measurement with stimuli from the current pool results in Spearman-Brown adjusted α's for both emotion methods that are notably below 0.70 (for more details, see Fig. S3 and Table S16 in the SOM). These results lend credence to the idea that verbal endorsements of anger versus disgust do not adequately distinguish between emotional responses to moral violations, but that such differentiation is possible (cf. Van der Eijk & Columbus, 2023). Further, given the strong correlation within emotion but across modalities – as well as similar inferences when only one of the two modalities is used – results suggest that assessing anger or disgust using only agreement with faces or only agreement with voices is defensible in contexts where only one option is available (e.g., if working with populations that cannot see or hear, or in contexts that do not allow for stimuli to either be seen or heard). Naturally, further methodological advances would be valuable, including further assessing the validity of these assessments in the populations sampled from here and in other populations.

4.4. Limitations and future research

The current project recruited participants from the United States and the Netherlands. Existing work suggests that the perception of moral emotion, especially disgust, varies across cultures (Han, Kollareth, & Russell, 2016; Kollareth & Russell, 2017), as does third-party punishment (Eriksson et al., 2021; House et al., 2020). Replication in different populations can inform the degree to which the distinct relations between anger and disgust versus direct and indirect aggression generalize to other cultures. Further, while we improved upon previous studies by using multiple moral violation vignettes and treating them as a random factor, generalization is limited to the types of violations we included. Replication using different methods can inform both the validity of the inference in this paper and the validity of different approaches to measuring disgust.

The current work shares another limitation present in other studies on this topic (e.g., Lopez et al., 2021; Molho et al., 2017): its reliance on self-reports of aggression motives. Other behavioral approaches, such as economic games, might yield different results. However, such methods do not necessarily distinguish between capturing the reputation management components of indirect aggression and the verbal and physical intervention components of direct aggression. Other behavioral paradigms like the hot sauce paradigm (Lieberman, Solomon, Greenberg, & McGregor, 1999) and the voodoo doll task (DeWall et al., 2013) similarly do not cleanly distinguish between direct and indirect aggression (Ritter & Eslea, 2005). Finally, in addition to limitations in terms of validity, single-item behavioral measures suffer from low reliability (Dang, King, & Inzlicht, 2020). Nevertheless, future work could take inspiration from longitudinal experience sampling or diary studies (e.g., Hofmann, Brandt, Wisneski, Rockenbach, & Skitka, 2018; Molho et al., 2020) to assess relations between the interpersonal value of moral violation targets and aggressive and emotional responses to transgressors. Such work could inform whether the sentiments assessed here correspond with behaviors, which might be constrained in ecologically-valid contexts.

Open practices

All materials, data, and analyses are available through the Open Science Framework (https://osf.io/36zxr/). Both studies were preregistered under this mentioned OSF project [links masked]; these preregistrations were completed prior to running the study and examining the data.

CRediT authorship contribution statement

Lei Fan: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Visualization, Writing – original draft. **Catherine Molho:** Conceptualization, Methodology, Resources, Writing – review & editing. **Tom R. Kupfer:**

Conceptualization, Methodology, Writing – review & editing. **Joshua M. Tybur:** Conceptualization, Investigation, Methodology, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare no competing interests. This project is funded by China Scholarship Council (201806990045).

Data availability

Data are available on OSF (https://osf.io/36zxr/).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2024.104597.

References

Alvarado, N. (1998). A reconsideration of the structure of the emotion lexicon. *Motivation and Emotion, 22*(4), 329–344. https://doi.org/10.1023/A:1021356424065

Archer, J., & Coyne, S. M. (2005). An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review, 9*(3), 212–230. https://doi.org/10.1207/s15327957pspr0903_2

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Barker, E. D., Tremblay, R. E., Nagin, D. S., Vitaro, F., & Lacourse, E. (2006). Development of male proactive and reactive physical aggression during adolescence. *Journal of Child Psychology and Psychiatry, 47*(8), 783–790. https://doi.org/10.1111/j.1469-7610.2006.01585.x

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review, 19*(4), 371–394. https://doi.org/10.1177/1088868314566683

Curtis, V., & Biran, A. (2001). Dirt, disgust, and disease: Is hygiene in our genes? *Perspectives in Biology and Medicine, 44*(1), 17–31. https://doi.org/10.1353/pbm.2001.0001

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass, 10*(2), 117–133. https://doi.org/10.1111/phc3.12192

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences, 24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007

De Vries, R. E., Lee, K., & Ashton, M. C. (2008). The dutch HEXACO personality inventory: Psychometric properties, self-other agreement, and relations with psychopathy among low and high acquaintanceship dyads. *Journal of Personality Assessment, 90*(2), 142–151. https://doi.org/10.1080/00223890701845195

Delton, A. W. (2010). *A psychological calculus for welfare tradeoffs*. Santa Barbara: University of California.

DeWall, C. N., Anderson, C. A., & Bushman, B. J. (2011). The general aggression model: Theoretical extensions to violence. *Psychology of Violence, 1*(3), 245–258. https://doi.org/10.1037/a0023842

DeWall, C. N., Finkel, E. J., Lambert, N. M., Slotter, E. B., Bodenhausen, G. V., Pond, R. S., Jr., … Fincham, F. D. (2013). The voodoo doll task: Introducing and validating a novel method for studying aggressive inclinations. *Aggressive Behavior, 39*(6), 419–439. https://doi.org/10.1002/ab.21496

Dinic, B. M., & Wertag, A. (2018). Effects of dark triad and HEXACO traits on reactive/proactive aggression: Exploring the gender differences. *Personality and Individual Differences, 123*, 44–49. https://doi.org/10.1016/j.paid.2017.11.003

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations, 19*(2), 152–168. https://doi.org/10.1177/1368430215583519

Eriksson, K., Strimling, P., Gelfand, M., Wu, J. H., Abernathy, J., Akotia, C. S., … Van Lange, P. A. M. (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature Communications, 12*, 1481. https://doi.org/10.1038/s41467-021-22955-x

Fan, L., Molho, C., Kupfer, T. R., Sauter, D. A., & Tybur, J. M. (2023). Beyond outrage: Observers anticipate different behaviors from expressors of anger versus disgust. *Social Psychological and Personality Science, 0*(0). https://doi.org/10.1177/19485506231176954

Fan, L., & Tybur, J. M. (2021). Emotional endorsement measurement of anger and moral disgust: A combination of facial and non-verbal vocal expressions. In *Kurt Lewin Instituut conference 2021*. Online.

Feinberg, M., Cheng, J. T., & Willer, R. (2012). Gossip as an effective and low-cost form of punishment. *Behavioral and Brain Sciences, 35*(1). https://doi.org/10.1017/S0140525X11001233

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–an analysis of articles published in 2015. *Journal of Experimental Social Psychology, 75*, 95–102. https://doi.org/10.1016/j.jesp.2017.11.008

Fischer, A. H., & Roseman, I. J. (2007). Beat them or ban them: The characteristics and social functions of anger and contempt. *Journal of Personality and Social Psychology, 93*(1), 103–115. https://doi.org/10.1037/0022-3514.93.1.103

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond purity: Moral disgust toward bad character. *Psychological Science, 28*(1), 80–91. https://doi.org/10.1177/0956797616673193

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Gutierrez, R., & Giner-Sorolla, R. (2007). Anger, disgust, and presumption of harm as reactions to taboo-breaking behaviors. *Emotion, 7*(4), 853–868. https://doi.org/10.1037/1528-3542.7.4.853

Han, D. H., Kollareth, D., & Russell, J. A. (2016). The words for disgust in English, Korean, and Malayalam question its homogeneity. *Journal of Language and Social Psychology, 35*(5), 569–588. https://doi.org/10.1177/0261927X15619199

Heerdink, M. W., Koning, L. F., van Doorn, E. A., & van Kleef, G. A. (2019). Emotions as guardians of group norms: Expressions of anger and disgust drive inferences about autonomy and purity violations. *Cognition & Emotion, 33*(3), 563–578. https://doi.org/10.1080/02699931.2018.1476324

Hildebrandt, A., Olderbak, S., & Wilhelm, O. (2015). Facial emotion expression, individual differences in. In J. D. Wright (Ed.) (2nd ed.)„ *Vol. 8. International encyclopedia of the social & behavioral sciences* (pp. 667–675). Oxford: Elsevier.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin, 44*(12), 1697–1711. https://doi.org/10.1177/0146167218775075

House, B. R., Kanngiesser, P., Bwrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., … Silk, J. B. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences, 287*(1925). https://doi.org/10.1098/rspb.2019.2794

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*(4), 719–737. https://doi.org/10.1037/a0022408

Jambon, M., & Smetana, J. G. (2018). Individual differences in prototypical moral and conventional judgments and Children's proactive and reactive aggression. *Child Development, 89*(4), 1343–1359. https://doi.org/10.1111/cdev.12757

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science, 17*(4), 283–286. https://doi.org/10.1111/j.1467-9280.2006.01699.x

Kirkpatrick, M., Delton, A. W., Robertson, T. E., & de Wit, H. (2015). Prosocial effects of MDMA: A measure of generosity. *Journal of Psychopharmacology, 29*(6), 661–668. https://doi.org/10.1177/0269881115573806

Knight, N. M., Dahlen, E. R., Bullock-Yowell, E., & Madson, M. B. (2018). The HEXACO model of personality and dark triad in relational aggression. *Personality and Individual Differences, 122*, 109–114. https://doi.org/10.1016/j.paid.2017.10.016

Kollareth, D., & Russell, J. A. (2017). The English word disgust has no exact translation in Hindi or Malayalam. *Cognition & Emotion, 31*(6), 1169–1180. https://doi.org/10.1080/02699931.2016.1202200

Kupfer, T. R., & Giner-Sorolla, R. (2017). Communicating moral motives: The social signaling function of disgust. *Social Psychological and Personality Science, 8*(6), 632–640. https://doi.org/10.1177/1948550616679236

Kupfer, T. R., & Giner-Sorolla, R. (2021). Reputation management as an alternative explanation for the "contagiousness" of immorality. *Evolution and Human Behavior, 42*(2), 130–139. https://doi.org/10.1016/j.evolhumbehav.2020.08.005

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud faces database. *Cognition & Emotion, 24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Lieberman, J. D., Solomon, S., Greenberg, J., & McGregor, H. A. (1999). A hot new way to measure aggression: Hot sauce allocation. *Aggressive Behavior, 25*(5), 331–348. https://doi.org/10.1002/(SICI)1098-2337(1999)25:5<331::AID-AB2>3.0.CO;2-1

Lopez, L. D., Moorman, K., Schneider, S., Baker, M. N., & Holbrook, C. (2021). Morality is relative: Anger, disgust, and aggression as contingent responses to sibling versus acquaintance harm. *Emotion, 21*(2), 376–390. https://doi.org/10.1037/emo0000707

Mathieson, L. C., & Crick, N. R. (2010). Reactive and proactive subtypes of relational and physical aggression in middle childhood: Links to concurrent and longitudinal adjustment. *School Psychology Review, 39*(4), 601–611. https://doi.org/10.1080/02796015.2010.12087745

Molho, C., Twardawski, M., & Fan, L. (2022). What motivates direct and indirect punishment? Extending the "intuitive retributivism" hypothesis. *Zeitschrift Fur Psychologie, 230*(2), 84–93. https://doi.org/10.1027/2151-2604/a000455

Molho, C., Tybur, J. M., Guler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological Science, 28*(5), 609–619. https://doi.org/10.1177/0956797617692000

Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications, 11*(1), 1–9. https://doi.org/10.1038/s41467-020-17286-2

Molho, C., & Wu, J. H. (2021). Direct punishment and indirect reputation-based tactics to intervene against offences. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 376*(1838). https://doi.org/10.1098/rstb.2020.0289

Nabi, R. L. (2002). The theoretical versus the lay meaning of disgust: Implications for emotion research. *Cognition & Emotion, 16*(5), 695–703. https://doi.org/10.1080/02699930143000437

Ocampo, D., Sullivan, J., Dayer, A., Palka, E., Betschart, N., & Holbrook, C. (2022). Prosocial aggression tracks genetic relatedness distinctly from emotional closeness. *Emotion*. https://doi.org/10.1037/emo0001175

Pedersen, E. J. (2015). *A welfare interdependence approach to third-party punishment.* University of Miami.
Piazza, J., & Landy, J. F. (2020). Folk beliefs about the relationships anger and disgust have with moral disapproval. *Cognition and Emotion, 34*(2), 229–241. https://doi.org/10.1080/02699931.2019.1605977
Piazza, J., Landy, J. F., Chakroff, A., Young, L., & Wasserman, E. (2018). What disgust does and does not do for moral cognition. *The Moral Psychology of Disgust*, 53–81.
R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98–103. https://doi.org/10.1016/j.tree.2014.12.003
Ritter, D., & Eslea, M. (2005). Hot sauce, toy guns, and graffiti: A critical account of current laboratory aggression paradigms. *Aggressive Behavior, 31*(5), 407–419. https://doi.org/10.1002/ab.20066
Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science, 5*(2). https://doi.org/10.1177/25152459221095827
Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76*(4), 574–586. https://doi.org/10.1037/0022-3514.76.4.574
Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion, 11*(2), 233–240. https://doi.org/10.1037/a0022598
Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science, 24*(10), 2069–2078. https://doi.org/10.1177/0956797613486988
Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology, 63*(11), 2251–2272. https://doi.org/10.1080/17470211003721642
Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., Rascanu, R., Sugiyama, L., Cosmides, L., & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition, 168*, 110–128. https://doi.org/10.1016/j.cognition.2017.06.002
Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In *Morality and health* (pp. 119–169). Taylor & Frances/Routledge.
Smith, A., Pedersen, E. J., Forster, D. E., McCullough, M. E., & Lieberman, D. (2017). Cooperation: The roles of interpersonal value and gratitude. *Evolution and Human Behavior, 38*(6), 695–703. https://doi.org/10.1016/j.evolhumbehav.2017.08.003
Sunar, D., Cesur, S., Piyale, Z. E., Tepe, B., Biten, A. F., Hill, C. T., & Koc, Y. (2021). People respond with different moral emotions to violations in different relational models: A cross-cultural comparison. *Emotion, 21*(4), 693–706. https://doi.org/10.1037/emo0000736
Sznycer, D., Sell, A., & Dumont, A. (2022). How anger works. *Evolution and Human Behavior, 43*(2), 122–132. https://doi.org/10.1016/j.evolhumbehav.2021.11.007
Sznycer, D., Sell, A., & Lieberman, D. (2021). Forms and functions of the social emotions. *Current Directions in Psychological Science, 30*(4), 292–299. https://doi.org/10.1177/09637214211007451
Tooby, J., & Cosmides, L. (1996). Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. In W. G. Runciman, S. J. Maynard, & R. I. M. Dunbar (Eds.), *Evolution of social behavior patterns in primates and man* (pp. 119–143). Oxford, England: Oxford University Press.
Tybur, J. M., Lieberman, D., Fan, L., Kupfer, T. R., & de Vries, R. E. (2020a). Behavioral immune trade-offs: Interpersonal value relaxes social pathogen avoidance. *Psychological Science, 31*(10), 1211–1221. https://doi.org/10.1177/0956797620960011
Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology, 97*(1), 103–122. https://doi.org/10.1037/a0015474
Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review, 120*(1), 65–84. https://doi.org/10.1037/a0030778
Tybur, J. M., Molho, C., Cakmak, B., Cruz, T. D., Singh, G. D., & Zwicker, M. (2020b). Disgust, anger, and aggression: Further tests of the equivalence of moral emotions. *Collabra: Psychology, 6*(1), 34. https://doi.org/10.1525/collabra.349
Van der Eijk, F., & Columbus, S. (2023). Expressions of moral disgust reflect both disgust and anger. *Cognition and Emotion, 37*(3), 1–16. https://doi.org/10.1080/02699931.2023.2183179
Veenstra, L., Bushman, B. J., & Koole, S. L. (2018). The facts on the furious: A brief review of the psychology of trait anger. *Current Opinion in Psychology, 19*, 98–103. https://doi.org/10.1016/j.copsyc.2017.03.014
Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17*(2), 267–295. https://doi.org/10.1037/emo0000226
Widen, S. C., & Russell, J. A. (2008). Children's and adults' understanding of the "disgust face. *Cognition and Emotion, 22*(8), 1513–1541. https://doi.org/10.1080/02699930801906744
Wu, J. H., Balliet, D., & Van Lange, P. A. M. (2016). Reputation, gossip, and human cooperation. *Social and Personality Psychology Compass, 10*(6), 350–364. https://doi.org/10.1111/spc3.12255
Wyckoff, J. P. (2016). Aggression and emotion: Anger, not general negative affect, predicts desire to aggress. *Personality and Individual Differences, 101*, 220–226. https://doi.org/10.1016/j.paid.2016.06.001
Yoshie, M., & Sauter, D. A. (2019). Cultural norms influence nonverbal emotion communication: Japanese vocalizations of socially disengaging emotions. *Emotion*. https://doi.org/10.1037/emo0000580