Journal of Experimental Psychology: General

© 2024 American Psychological Association
ISSN: 0096-3445

2024, Vol. 153, No. 6, 1551–1567
https://doi.org/10.1037/xge0001570

Affective Prediction Errors in Persistence and Escalation of Aggression

Marius C. Vollberg[1, 2] and Mina Cikara[3]
[1] Swiss Center for Affective Sciences, University of Geneva
[2] Department of Psychology, University of Amsterdam
[3] Department of Psychology, Harvard University

People generally empathize with others and find harm aversive. Yet aggression, for example, between groups, abounds. How do people learn to overcome this aversion in order to aggress? Many models of learning emphasize outcome prediction errors—deviations from expected outcomes in the environment—but aggression may also be fueled by affective prediction errors (affective PEs)—deviations from how we expect to feel. Across five preregistered online experiments that hold outcome prediction errors constant ($N = 4{,}607$), participants choosing aggressive or nonaggressive actions aggressed more against disliked group members and often escalated or persisted in taking actions that felt better than expected (positive affective PE), especially when those actions were aggressive. Crucially, inducing incidental empathy toward the group of the target rendered affective PE signals sensitive to group identification—participants escalated aggression that felt better than expected relatively less toward liked versus disliked group members. That said, affective PEs did not always add explanatory power beyond levels of postoutcome affect alone; we discuss the importance and implications of these results. In summary, we reveal affective PE integration as a candidate algorithm facilitating exceptions to harm aversion in intergroup conflict. More broadly, we highlight for affective science and decision-making researchers the necessity of appropriately testing separable components of affective signals in predicting subsequent behavior.

Public Significance Statement

How do people learn to choose one action over another? Many models focus on learning from prediction errors—differences between expectations and outcomes. Research on prediction errors has historically focused on differences between expectations and outcomes in one's environment (e.g., winning more money than expected from a slot machine), but people ultimately choose based on the subjective value of those outcomes, which may be approximated via self-reported affect, or feelings. We wanted to understand whether feeling better than expected about an action informs what we do, especially in contexts without (informative) outcome prediction errors. In short, the answer was yes, but it is complicated. Affective prediction errors (affective PEs) predicted subsequent aggressive choices but not always over and above just how good participants felt after taking the aggressive action. This is important because it highlights a core distinction between prediction errors about the world and prediction errors about the self. Moreover, incidentally giving participants the opportunity to empathize with their target's group attenuated the relationship between affective PEs and aggression escalation, but only toward liked targets. Our results show that, in the test case of aggression, choices can track with affective PEs and that incidentally induced intergroup affect may influence this association.

Keywords: harm aversion, learning, affective prediction errors, intergroup aggression

Supplemental materials: https://doi.org/10.1037/xge0001570.supp

This article was published Online First May 2, 2024.

Lotte Frederike Van Dillen served as action editor.

Marius C. Vollberg https://orcid.org/0000-0003-2301-3552

This research was supported by a National Science Foundation Grant (1551559 awarded to Mina Cikara) and a Swiss National Science Foundation fellowship (P500PS_210892 awarded to Marius C. Vollberg). We would like to thank Daniel Gilbert, Todd Hare, Liz Phelps, and David Sander for their helpful inputs at various stages of the project.

Parts of the research presented in this article have been presented at the Association for Psychological Science's International Convention of Psychological Science and the Society for Personality and Social Psychology's Annual Convention. All the data, materials, and analyses are available on Open Science Framework under the following link: https://osf.io/fp9hk/?view_only=4c9131ef5927417aa9e255ea40881820.

Marius C. Vollberg served as lead for conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing–original draft, and writing–review and editing. Mina Cikara served as lead for funding acquisition, project administration, resources, and and served in a supporting role for investigation, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Marius C. Vollberg, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, Netherlands. Email: mariusvollberg@gmail.com

There is mounting empirical evidence that people generally find harming others aversive; for example, they would rather harm themselves than others and exhibit physiological arousal when merely simulating aggression (Crockett et al., 2014; Cushman et al., 2012). Although people tend to avoid aggression and, historically speaking, commit less and less of it (Pinker, 2011), intergroup conflict remains a threat to human development, the world over (Woolf & Hulsizer, 2004).

If people are so averse to aggressing against others, how do they overcome this aversion to engage in intergroup aggression? Past literature has focused largely on identifying the situational features that foment aggression without providing an explanation for how generally averse behaviors might transition from unsavory to appetitive (Cikara & Van Bavel, 2014; Darley & Latane, 1968; Hogg, 1993; Zhong et al., 2010; Zimbardo, 1995; see, however, Bandura, 1976; Chester, 2017; Cikara, 2015; Fiske & Rai, 2014; we return to some of this work shortly). Here, we explore a potential mechanism to help answer this question: learning one's own aggression-related preferences via affective experience.

Defining Affective Prediction Errors (Affective PEs)

Changing one's preferences or beliefs as a function of experience is often characterized through the lens of reward learning. In a canonical example task, people make predictions (e.g., to win $5 in a lottery), observe outcomes (e.g., actually win $10), and update their predictions based on the differences between expectations and outcomes, referred to as prediction errors. Largely operationalized via material outcomes in the environment (e.g., dollars, mL of juice) as units of utility, such "objective" value-based prediction errors have proved a powerful algorithmic tool guiding adaptive behavior in humans and machines (Daw et al., 2011; Mitsuko et al., 2017; Schultz et al., 1997). But we can also predict how we will feel about an outcome (affective prediction or forecast) and end up actually feeling differently (affective PE), and this subjective signal may track with subsequent choices.

We propose that people may learn to aggress via experience and that affective PEs may act as a critical learning signal, which can be approximated via self-report. The central role of affect is widely recognized in learning and decision-making research (Charpentier et al., 2016; Engelmann et al., 2019; Garrett et al., 2016; Knutson & Greer, 2008; Mellers & Mcgraw, 2001; Mellers et al., 1999; Phelps et al., 2014; Sander & Nummenmaa, 2021). Investigations of prediction errors, more specifically, have started to incorporate affect, too (Rutledge et al., 2014). But in these cases, affect is often reduced to a downstream consequence of outcome prediction errors (e.g., obtaining more money than expected makes people happier). Although prediction errors at the level of affect itself have already been documented widely in the affective forecasting literature (Nielsen et al., 2008; Wilson & Gilbert, 2005), they are less often used to predict subsequent behavior (Collins & Shenhav, 2021). A recent study sought to address this gap in the context of economic games, separating emotion prediction errors (measuring both valence and arousal) and outcome prediction errors, suggesting that they might independently and systematically contribute to choice behavior (Heffner et al., 2021). That there is explanatory power of affective PEs above and beyond outcome prediction errors has at least two implications that complement traditional reward learning frameworks (see Vollberg & Sander, 2024 for a detailed discussion of how affective PEs relate to traditional reward learning frameworks). First, verbally reported affect can be used to approximate subjective value that reflects variations in an organism's internal and external states (Juechems & Summerfield, 2019). Second, individuals may learn not just from mispredicting the world (i.e., outcomes in the environment) but also from mispredicting themselves (i.e., their own subjective value states approximated via affect).

The literature on reward learning and affective forecasting conflict in their implied importance of affective PEs. From a reward learning perspective, prediction errors are a core input for updating; their affective counterpart would be expected to guide behavior, too (all else being equal). From an affective forecasting perspective, however, affective traces are quick to fade and expectations may not even be relevant at all for how we feel and act after we experience an outcome (Golub et al., 2009). In principle, because affect (as opposed to outcomes in the environment) originates in the experiencer, prediction errors about affect may follow different rules than prediction errors about outcomes, despite their superficial similarities (Vollberg & Sander, 2024). As we test our novel framework in the context of aggression, we will thus make sure to keep these alternatives in mind empirically. Specifically, we will test whether affective PEs explain variance in behavior above and beyond just postoutcome affect. A priori, there is reason to believe that affective PEs are likely to factor into decision making about aggression precisely because aggressive behaviors are widely associated with negative rather than rewarding consequences (Cushman et al., 2012) and yet there are contexts in which they will feel good (e.g., because they fulfill a goal, such as winning a boxing match or disarming a violent perpetrator).

Affect as Prediction Error in Intergroup Aggression

Humans tend to categorize themselves and others into groups, which in turn determine with whom we cooperate and compete (Allport, 1954; Tajfel & Turner, 1979). We tend to cooperate with those we deem part of our group and to aggress against those who we perceive to pose a threat (Chang et al., 2016). We therefore predict that people will aggress more against threatening outgroup than ingroup members.

The affective footprints accompanying this intergroup bias in behavior have been studied across contexts and measures, revealing that reward prediction error-related brain regions also correlate with pleasure in response to disliked others' misfortunes or pain. For example, pleasure that participants report in response to watching outgroup failure correlates with ventral striatum (VS) activity (Cikara et al., 2011); VS activity is in turn associated with subsequent aggression (Chester & DeWall, 2016). Studies in rodents, too, document dopamine spikes within VS after attacking an intruder (Van Erp & Miczek, 2000). The VS is thought to serve reinforcement learning by encoding rewarding events to maximize likelihood of reward in the future (Bartra et al., 2013). By extension, this evidence suggests that, in some contexts, aggression carries intrinsic value just like other reinforcers (Chester, 2017; Cikara, 2015). But how does aggression acquire that positive value? We hypothesize this reinforcing quality to manifest in affect, such that the degree to which aggression feels better than expected comes with an increase in both the likelihood of continued aggression (i.e., "persistence") and the extent of that aggression (i.e., "escalation").

Although affective PEs might be predictive of behavior in general, the specific type of behavior (e.g., aggression) might moderate this relationship. From a functional perspective, if the computational goal is to behave prosocially most of the time while allowing one to switch to aggression when necessary (e.g., in the presence of a threatening outgroup member), a plausible algorithmic approach would be to generally assign a negative value to aggression that can be quickly overwritten given exceptional circumstances. From this perspective, aggression feeling better than expected (positive affective PEs), for example, through the absence of punishment, could be an algorithmic tool to signal "in this situation, aggression is an appropriate behavior." Compared to neutral behavior, this signal could be encoded via increases in the size or the weight of the prediction error. That is, aggressive acts may feel better than expected compared to nonaggressive acts (i.e., difference in prediction error size), or, alternatively, aggressive acts feeling better than expected may impact subsequent behavior more compared to nonaggressive acts feeling better than expected, regardless of prediction error size (i.e., difference in prediction error weight). Although both paths could ultimately result in the same value signal (Levy & Glimcher, 2012), we cannot know a priori which of them better characterizes affective PEs in aggression. We thus predict affective PEs to be increased either in size or in terms of their predictive weight in aggressive compared to neutral behavior.

There is also likely interindividual variation in the association between affective PEs and aggression toward liked versus disliked targets at baseline. Probing systematic variability of these associations requires experimental manipulation. We do this via emotion induction prior to the task. When we observe others' pain or misfortunes, we tend to feel empathy for ingroup members but pleasure or schadenfreude for competitive outgroup members (Cikara et al., 2014), which, as we have already noted, may facilitate aggression. Capitalizing on these liked-target/empathy and disliked-target/schadenfreude associations, we test whether activating intergroup emotions (vs. a neutral emotion control) prior to the main task changes how people escalate aggression that felt better than expected depending on the target's group (Cikara, 2015). We predict that inducing counter-empathy for disliked groups will strengthen the association between affective PEs and escalation of aggression toward disliked relative to liked targets (relative to a neutral induction condition).

Hypotheses and Overview of Experiments

In general, we hypothesized that competitive intergroup aggression can be understood in part as the result of affective PEs. We predicted that across competitive settings participants would choose to take aggressive actions more toward disliked than neutral groups or ingroups (Hypothesis 1 [H1]). We predicted that, on average, affective PEs would be larger for aggressive compared to neutral behavior (Hypothesis 2 [H2]). We further predicted the persistence (Hypothesis 3 [H3]) and escalation of an action (Hypothesis 4 [H4]) to increase with the degree to which an action felt better than expected across both neutral and aggressive actions (i.e., the size of the affective PE; Hypothesis 3a [H3a] and Hypothesis 4a [H4a]). In case affective PEs show no mean size difference between aggressive and neutral actions (i.e., no evidence for H2), we predicted increased weights of affective PEs on subsequent persistence and escalation for aggressive compared to neutral actions (Hypothesis 3b [H3b] and Hypothesis 4b [H4b]). Lastly, we hypothesized the association between affective PEs and aggression would be moderated by inducing intergroup emotion prior to the task. Specifically, we predicted that prior exposure to others' suffering would yield a relatively stronger association between affective PEs and the escalation of aggression against disliked relative to liked groups (Hypothesis 5 [H5]). We test H1, H3a, H3b, and H4a across all five experiments; H2 and H4b in Experiments 1 and 1b; and H5 in Experiments 3 and 3b.

Experiment 1

Method

Overview

We developed a multitrial game in which we told participants they were playing with another person (target; see Figure 1 for an overview of the experimental design). On each of 20 trials we recorded participants' affective predictions about their behavior (i.e., how they thought they would feel about taking an other-neutral or other-harmful action), the behavior itself (i.e., whether they chose the neutral or aggressive action), and affective outcomes (i.e., how they actually felt about what they had just done). Both actions were equally beneficial for the participant.

We operationalized aggression as stealing virtual points. Although harm and aggression are often used interchangeably, referring to stealing as aggression relates to the underlying "Point Subtraction Aggression Paradigm" (PSAP; Cherek et al., 1997) from which our task is adapted along with the view of aggression as a molecular constituent of harm. Specifically, on each trial, participants were asked to choose between stealing points from the target and earning points for themselves. They could choose the policy (earning or stealing) and the amount (1–8 points). Policy persistence thus refers to choosing the same action on two subsequent trials, while policy escalation refers to increasing the chosen amount within policies. Crucially, participants always received the outcome they chose, which fixed outcome prediction errors to zero. We told participants that the target could also steal from them, but that stealing on every trial implied mutual destruction. Seeing that we did not want to introduce retribution as another motive, we informed participants that we would not tell them whether the target was stealing until the end (we never ended up providing such information at all). We made the task equally competitive across all conditions by increasing the bonus payment for whichever player had more points at the end of the task. We specifically manipulated the intergroup context by randomly assigning participants to play with a target who belonged to one of two groups. This general structure forms the basis for all five experiments, but outcome contingencies, group identities, and an emotion induction were modified or introduced to test specific hypotheses and increase generalizability (see Table 1 in the online supplemental materials).

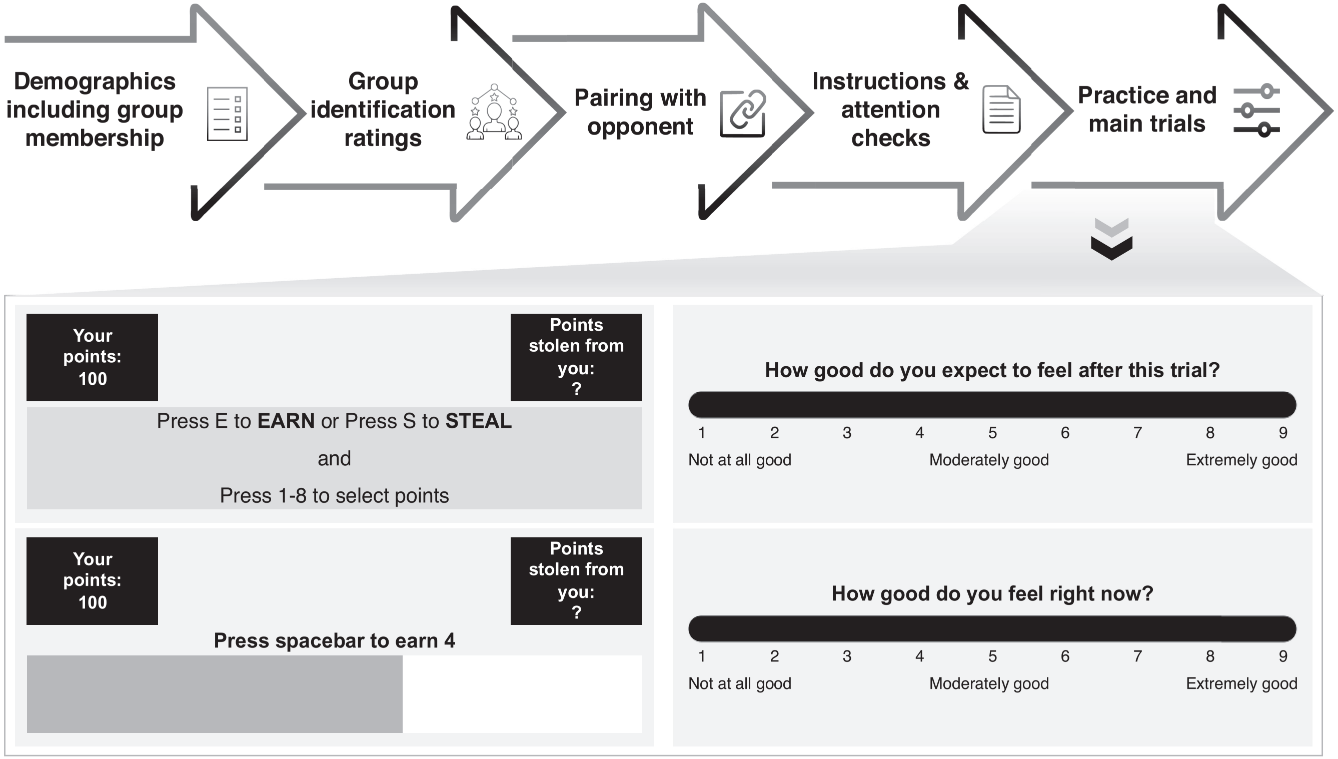
Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All the data, analysis code, and research materials are available at https://osf.io/fp9hk/?view_only=4c9131ef5927417aa9e255ea40881820. The preregistration for this experiment is available at https://osf.io/2dmtw/?view_only=4952a20353ea4730bc80b3070ad57ccd.

Participants

Our experimental design let participants choose freely, and some of our hypotheses require participants to have chosen the same policy on at least two consecutive trials. Consequently, the number of

Figure 1
Overview of Experimental Design


Demographics including group membership
Group identification ratings
Pairing with opponent
Instructions & attention checks
Practice and main trials
Your points: 100
Points stolen from you: ?
Press E to EARN or Press S to STEAL
and
Press 1-8 to select points
How good do you expect to feel after this trial?
1 2 3 4 5 6 7 8 9
Not at all good
Moderately good
Extremely good
Your points: 100
Points stolen from you: ?
Press spacebar to earn 4
How good do you feel right now?
1 2 3 4 5 6 7 8 9
Not at all good
Moderately good
Extremely good

Note. Participants were ostensibly matched with another player (target) based on their demographic information and were informed about the target's age, gender, and group membership, with the latter being the only manipulated variable. Following instructions and attention checks, participants began the main block consisting of a modified PSAP (see the Method section). Participants could choose between aggressive (stealing) and neutral (earning) strategies, as well as the corresponding amount. They were then asked how good they expected to feel after a given trial before completing a button-pressing task under generous time pressure to obtain their chosen outcome. At the end of each trial, participants were asked to rate their postoutcome affect. Participants' own scores were updated based on their choices, but the amount stolen by the ostensible target was not shown to participants (i.e., a question mark appeared throughout the task and participants received no information about the target's actions until they were debriefed. PSAP = Point Subtraction Aggression Paradigm.

usable participants may vary by analysis (sample sizes specified in all figures and analyses). To account for this variability, we based our sample size on conservative simulations of the effect most affected by this fact: the association of policy escalation within stealing (increases in the amount stolen from $t-1$ to t) and affective PEs. The simulations revealed that around 150 participants would be necessary to detect the association of the magnitude found in pilot data with 80% power. Pilot data suggested that about a quarter of participants would qualify for the analysis contingent on consecutive stealing, which corresponds to a required 600 participants remaining after preregistered exclusions (attention check, meaningful responses; see below). To account for noise in the pilot data and to increase the estimated power to 90%, we collected 1,000 submissions, aiming for 850+ usable participants. All participants were recruited in the United States via Prolific, paid at the rate of roughly $10 per hour before bonus, aged 18 years or older, and provided informed consent to participate in the experiment, which was approved by the university's Institutional Review Board.

Out of 1,000 complete submissions, nine participants were excluded because they stated to have experienced more than zero fatal heart attacks in their lives, 61 participants were excluded because they averaged less than half of the required (10) button presses, and 20 participants were excluded because they gave invalid keyboard input. Following exclusions, this yielded a final sample of 910 participants (449 women, 436 men, 25 other, $M_{\text{age}} = 32.9$, age range = 18–77, $SD = 11.35$, 714 Democrat).

Manipulations and Measures

Demographics. Participants were asked to report their gender, ("With which gender do you most identify?"; response options: "Man," "Woman," "Nonbinary," "Another gender not listed here," and "I do not wish to provide this information"), political affiliation ("With which political party do you most identify?"; response options: "Democrat" and "Republican"), and age (dropdown list from 18 to 100).

Identification Score. Participants were presented with three questions asking how much they "like," "value," and "feel connected" to Republicans and Democrats. All three questions were presented on the same page with response options on a continuous scale from *not at all* over *somewhat* to *very much* (internally coded from 0 to 100; each party appeared on its own page in randomized order). Responses converged across these three items (Cronbach's $\alpha > .9$ in all experiments). We thus averaged responses for both groups, respectively. Participants were matched and lead to believe to be interacting with only one other person, representing only one party. Contrary to several intergroup studies where participants interact with members of both groups, we thus used the mean identification ratings for that target group rather than the bias (difference between target group and the other group) for our analyses.

Adapted PSAP. There are several ways to measure aggression. One validated approach is to let participants choose between actions that do versus do not administer an aversive stimulus to a target. In the PSAP (Cherek et al., 1997), the aversive stimulus consists of subtracting points from the target. Here we used a variant of the PSAP with two options: earning for oneself, or stealing from the target ("E" and "S" buttons, respectively). Within those choices, participants could further indicate how much they would like to earn or steal between 1 and 8 points (numbers 1–8 on the keyboard). On a separate screen, to realize their preference (i.e., actually earn or steal), participants had to press the space bar 10 times; the button presses were visually represented by a progress bar that filled up correspondingly. Depending on whether the button was pressed a sufficient number of times within a fixed time window of 3 s, either a green hook or a red cross would appear for feedback. The initial score was 100 points, and while the participant's own score was shown and updated on each trial, the target's score and stealing behavior were not shown to participants at any point. The reasoning behind not providing such information or introducing actual interactions with another player was to minimize explicit motivation to retaliate.

Predicted Affect. Measuring affective PEs requires affective predictions and affective outcomes. We measured predicted affect in between participants' PSAP choices and their actual attempts to complete the task by asking them: "How good do you expect to feel after this trial?" on a continuous scale from *not at all good* over *moderately good* to *extremely good* (internally coded as 0–100).

Postoutcome Affect. Complementing predicted affect, we measured postoutcome affect after participants' attempts to complete the task by asking them: "How good do you feel right now?" on a continuous scale from *not at all good* over *moderately good* to *extremely good* (internally coded as 0–100).

Comprehension Checks. We used comprehension checks to ensure that participants understood who they were stealing from, what group the target belonged to, and that the target could also steal from them. These checks were administered in the form of questions with multiple response options (one, two, and two distractors, respectively) on three separate screens. The first screen asked: "When you choose to steal, are you stealing from the experimenters or from the other player?." The second asked: "Which political affiliation is the other player?." Lastly, the third screen asked: "Who can choose to steal?." Participants received negative feedback upon choosing the wrong response and could only proceed after choosing the correct response.

Attention Check. We included a general attention check where participants were asked to state how many fatal heart attacks they had experienced in their lives.

Procedure

Following consent and the attention checks, participants received general instructions about the task. They were told that they are about to do a task in which they should try to gain as many points as possible. To emphasize competition, participants were told that their bonus would be reduced for every 10 points that their target gained more than them. Participants were then asked to answer a set of questions to be paired with another player before being presented with a screen on which they were asked to report their gender, political affiliation, and age. On the subsequent two screens, they were asked to provide the three identification ratings ("like," "value," and "feel connected to"; see the Manipulations and Measures section) about Democrats and Republicans, respectively (in randomized order). Following a short loading animation during which the pairing with another player was ostensibly processed, participants were then informed that they had been matched with a 30-year-old male who was either Republican or Democrat (between-participant condition). Age and gender were meant to roughly reflect the modal participant in online studies.

Having been paired with the target, participants now received more extensive instructions about the task. First, they were reminded that they would be playing against the person they had been matched with. They were then told that they would start out with 100 points, that they would be able to choose between earning and stealing on each trial, and that those options were also available to the target. They were informed about the specific requirements of the task in which they would have to press the space bar "approximately 10 times, as quickly as you can, in order to complete the round successfully," and that points are only allocated if they actually complete the task. To further clarify the mutual destruction that would result from escalated aggression, participants were also told that it is possible to obtain a negative score. Lastly, participants also received instructions about the emotion ratings that would follow each PSAP choice and task attempt, highlighting the difference between expected and postoutcome emotions.

Next, participants were presented with the three comprehension checks recapitulating the key features of the experiment before completing two practice trials. All trials consisted of: the PSAP choice, predicted affect rating, PSAP task, and postoutcome affect rating. The time window for the PSAP task was extended from 3 to 5 s during practice trials. Participants then completed 20 main trials. After those 20 trials and before being debriefed, participants were asked to provide a second round of identification ratings. This was included for exploratory analyses, but, as preregistered, only the initial identification scores were used for the present analyses.

Analysis

The central analyses assess the effect of affective PEs. Affective PEs were calculated by taking the difference between ratings of postoutcome affect and ratings of predicted affect on every trial:

$$\text{aPE}_t = \text{postoutcome affect}_t - \text{predicted affect}_t. \tag{1}$$

Per definition, affective PEs differ from outcome prediction errors in that the object of prediction is subjective and based on introspection. Outcome prediction errors can serve to incrementally adjust expectations toward a ground truth based on outcomes in the environment. Affective PEs, on the other hand, do not have such a ground truth to converge on. Correspondingly, research on affective forecasting (Meyvis et al., 2010) and recent work comparing outcome prediction errors to emotion prediction errors (Heffner et al., 2021) suggests a more transient signal that does not cohere across time as precisely as outcome prediction errors do. Reflecting this temporal feature, each affective PE was used to predict behavior only from one trial to the next (see, however, the online supplemental materials for analyses of affective PEs across trials). Specifically, affective PEs at $t - 1$ were used to predict the chosen policy at t (H3a and H3b), as well as the amount chosen within policy at t (H4a, H4b, and H5). The remaining hypotheses more straightforwardly predict the probability of stealing as a function of target identification (H1), as well as the absolute size of affective PEs as a function of the chosen policy (i.e., aggressive vs. neutral; H2).

How can we interpret responses across participants given that the underlying scale is arbitrary and might be used differently by different participants? If we imagine two different people feeling the exact same way, one might still rate their happiness at 40/100, while the other rates it at 60/100, which corresponds to a misleading difference in means. Moreover, the two hypothetical participants may differ in how much they systematically over- or underestimate how they will feel. Here, we were interested in how prediction errors relative to a given participant's standards predict behavior. Given that we collect multiple responses from each participant, we can treat the data accordingly by standardizing it within participants: affective PEs are transformed by subtracting a given participant's mean affective PE from the affective PE of each of their 20 trials and subsequently dividing each of those resulting values by that same participant's standard deviation in affective PEs. One unavoidable consequence of standardizing in this fashion is that participants without any variability in their responses across trials whatsoever cannot be subjected to standardization and are thus effectively excluded from relevant analyses (we specify sample sizes in all figures and analyses to show variability due to standardization or monotonous choice patterns).

We also transformed the target group identification ratings. Because participants did not provide repeated measures on this variable, standardization was implemented between participants. The variable "target identification" used in all analyses thus refers to a given participant's mean across how much they like, value, and feel connected to the group of the target they were matched with, relative to other participants, expressed in standard deviations.

All analyses were conducted using linear regression models accounting for the distribution of the response variable and repeated measures. Specifically, we used logistic regression for dichotomous outcomes—such as aggressive versus nonaggressive policies, or policy switching versus policy persistence, from $t - 1$ to t—implemented with the "glmer" function in the "lme4" package (Bates et al., 2015). Continuous outcomes—such as increases in the amount chosen from $t - 1$ to t—were analyzed using linear regression using the "lmer" function from the same package. Both approaches included random effects (intercept and slope depending on the model fit; see below) to cluster error related to repeated measures, such as participant ID, trial number, and policy type as applicable. For each analysis, we selected the model with the lowest Akaike information criterion out of a range of models up to the maximal random effects structure that still allowed model convergence; in the case of equal Akaike information criterion values, the model closer to the maximal specification was selected. We assessed normality in fitted residuals using the DHARMa package (Hartig, 2018). This was particularly important to ensure that deviations from normality in the underlying data (e.g., multimodality) did not violate the assumptions of the regression models we used. Partial eta squared effect sizes (η^2) were calculated by taking the estimated fixed effects and dividing them by the square root of the sum of random effect variances of the corresponding model (Brysbaert & Stevens, 2018).

We will now outline what our analyses specifically translate to with respect to our preregistered hypotheses (the code implementing these all other analyses can be found on the Open Science Framework repository linked above).

0. Do people identify more with their ingroup?

$$\text{identification} \sim b_0 + b_1\alpha + \varepsilon_s + \varepsilon_t, \tag{2}$$

where b represents regression weights of the following fixed effects: intercept (b_0) and ingroup/outgroup (α). The maximal random effects structure included a random intercept for repeated measurements per participant (ε_s).

1. Do people steal more from people they don't like?

$$\text{amount stolen} \sim b_0 + b_1\tau + \varepsilon_s + \varepsilon_t, \tag{3}$$

where b represents regression weights of the following fixed effects: intercept (b_0) and target identification (τ). The maximal random effects structure included random effects for repeated measurements per participant (intercept only; ε_s) and trial number (slope in maximal model; ε_t).

2. Are affective PEs higher for stealing compared to earning?

$$\text{aPE} \sim b_0 + b_1\delta + \varepsilon_s + \varepsilon_t, \tag{4}$$

where b represents regression weights of the following fixed effects: intercept (b_0) and the policy type (earning or stealing; δ). The maximal random effects structure included random effects for repeated measurements per participant (intercept only; ε_s) and trial number (slope in maximal model; ε_t).

3. Are affective PEs at $t - 1$ positively associated with choosing the same policy at t (policy persistence)?

$$\text{p(stay)} \sim b_0 + b_1\alpha + \varepsilon_s + \varepsilon_t, \tag{5}$$

where b represents regression weights of the following fixed effects: intercept (b_0) and affective PE at $t - 1$ (α). The maximal random effects structure included random effects for repeated measurements per participant (ε_s), and trial number (slope or intercept; ε_t). For H3b, we also included the policy type (δ) as a main effect in interaction with (α), as well as the corresponding random effect for policy type (slope in maximal model; ε_u).

4. Are affective PEs at $t - 1$ positively associated with increases in the amount chosen at t (conditioning on same policy being chosen; policy escalation)?

$$\Delta\ \text{amount} \sim b_0 + b_1\alpha + \varepsilon_s + \varepsilon_t, \tag{6}$$

where b represents regression weights of the following fixed effects: intercept (b_0) and affective PE at $t - 1$ (α). The maximal random effects structure included random effects for repeated measurements per participant (intercept only; ε_s), trial number (slope in maximal model; ε_t). For H4b, we also included the policy type (δ) as a main effect in interaction with (α).

Exploratory Analyses. An increasing shift toward computational modeling in cognitive and psychological science has introduced new approaches to data analysis in experimental psychology. Despite the promise of this development, varying standards in methodological transparency can sometimes obscure what aspects of an analysis framework are necessary versus "just" ideal. Seeing that we are presenting a new, potentially generative approach, we sought to probe and make transparent its robustness to varying analytic strategies. To this end, we provide an overview of whether and how our results change when using unstandardized and between-participant standardized affective PEs instead of our within-participant standardization, as well as when adjusting for demographic covariates.

Perhaps the most theoretically important robustness check focuses on the predictive power of how good people felt after each trial (i.e., examining the effect of affective PEs above and beyond postoutcome affect; see Table 2 in the online supplemental materials). Controlling for postoutcome affect is important because it allows us to address the tension between the affective forecasting and reward learning literature noted in the introduction. We thus report an overview of these analyses at the end of each study. In principle, if we want to parsimoniously predict future behavior, we need to know whether "feeling better/worse than expected" explains sufficient variance above and beyond merely "feeling good/bad" after experiencing an outcome. But why not also account for "having expected to feel good/bad"? Although the strongest candidate for predicting behavior would appear to be the affective measure closest to the outcome, we also include analyses controlling for predicted affect (see Table 2 in the online supplemental materials). Note that we follow the precedent of Heffner et al. (2021) who assessed the relative predictiveness of composite prediction errors by including predicted affect in the same model jointly with affective PEs. However, this approach comes with the risk of suppression effects (indeed, the signs of associations often did flip in our analyses when moving from separate to joint estimation; see Table 2 in the online supplemental materials). We thus additionally provide another set of analyses where models were estimated separately before comparing the respective model fits. Note that, in order to do this, affective responses were no longer standardized within participants, ensuring that the same observations are included in both models (i.e., avoiding imbalance introduced by failed standardization of invariant prediction errors within a given person).

Additional analyses assess the evolution of affective PEs over time and the role of group identification in the association between affective PEs and behavior (see also the online supplemental materials). Note that these exploratory analyses are neither preregistered nor in conflict with the main findings presented in this article. Of note, group identification showed no evidence for an interaction with the association between affective PEs and behavior in pilot data. We therefore deemphasized this variable (see the online supplemental materials for converging evidence), except for Experiments 3 and 3b, in which we induced intergroup emotion and thus expected group identification to play a role.

Results

Experiment 1 tested H1, H2, H3, and H4. Specifically, we tested whether participants would steal more from targets of disliked groups using political parties (Democrats vs. Republicans; H1). Note that, for inferential analyses, we used a more fine-grained, continuous measure of target identification—how much participants like, value, and feel connected to a given target—in lieu of dichotomous group membership to account for variability in intergroup animus. We predicted that mean affective PEs would be increased for aggressive compared to neutral actions (H2). We further predicted that affective PEs at $t - 1$ are positively associated with both choosing the same action (policy persistence; H3a) and increases in the amount at t (policy escalation; H4a). In case affective PEs show no difference by policy type (i.e., no evidence for H2), we also predicted these associations to be stronger for aggressive relative to neutral actions (H3b and H4b, respectively). Hypotheses involving persistence and escalation are assessed in participants who chose the same relevant behavior twice in a row at least once; corresponding analyses thus subset on participants who meet this criterion, causing sample sizes to vary accordingly (as specified in all figures and results).

Manipulation Check and H1: Group Bias

Consistent with general harm aversion, people chose earning more often than stealing (29.6% of all choices). As predicted, however, the intergroup manipulation yielded bias in both attitudes (manipulation check) and stealing (supporting H1). People identified less with outgroup ($M_{\text{outgroup}} = 22.6$, $SD = 21.3$) compared to ingroup ($M_{\text{ingroup}} = 58.7$, $SD = 25.8$) members, $b = -36.183$, 95% confidence interval (CI) [−38.152 to −34.214], $\eta^2 = -1.53$, $n = 910$, $p < .001$, and stole less from targets with whose group they identified more, $b = -0.254$, 95% CI [−0.437 to −0.071], $\eta^2 = -0.07$, $n = 910$, $p = .007$ (see Figure 2B; see the online supplemental materials for corresponding figures for Studies 1b, 2, 3, and 3b). (We did not expect an interaction with target group membership because the identification measure already accounts for this variable.)

H2: Affective PEs for Stealing Versus Earning

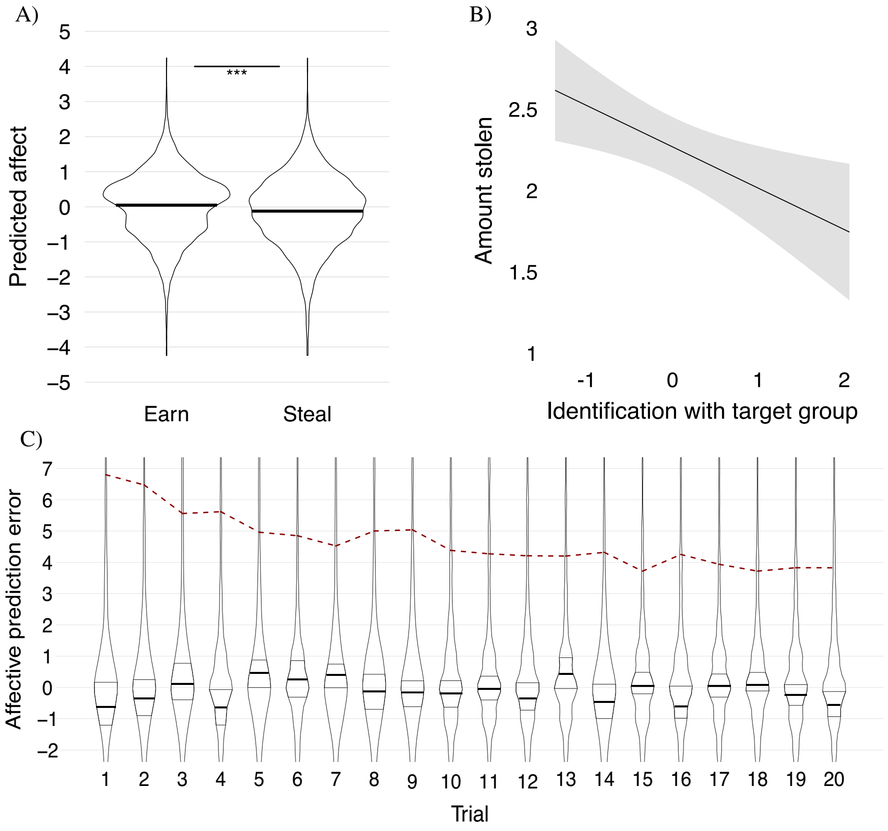
Counter to our hypothesis, participants did not report larger affective PEs following stealing trials compared to earning trials, $b = 0.015$, 95% CI [−0.015 to 0.046], $\eta^2 = 0.02$, $n = 901$, $p = .325$.

H3a and H3b: Policy Persistence

Counter to our hypotheses, affective PEs showed no association with subsequent policy persistence: neither across policies nor when comparing stealing to earning. The association between

Figure 2

Manipulation Checks and Descriptive Analyses in Experiment 1



Note. $N = 910$. (A) Distributions of within-participant standardized affective predictions (y axis) separated by policy (earning and stealing; x axis). Taking into account each participant's standards, participants expected stealing to feel worse than earning, $\beta = -0.169$, 95% CI [−0.197 to −0.141], $\eta^2 = -0.14$, $n = 896$, $p < .001$ (***). (B) Model estimate of the relationship between standardized identification with the target's group on the x axis and the absolute amount stolen from them on the y axis; participants stole less from members of groups that they liked, valued, and felt connected to more. Shaded area reflects 95% CI. (C) Distributions of signed affective PEs (y axis) across trials (x axis) with mean absolute (unsigned) values shown in red. While signed values average around zero, showing no overall imbalance between over- and underestimations, $b = -0.125$, 95% CI [−0.432 to 0.182], $\eta^2 < -0.01$, $n = 910$, $p = .424$, absolute (unsigned) values revealed decreasing prediction errors over time, $b = -0.133$, 95% CI [−0.163 to −0.103], $\eta^2 = -0.01$, $n = 910$, $p < .001$. Subsequent experiments show similar patterns (see the online supplemental materials). CI = confidence interval; affective PE = affective prediction error. See the online article for the color version of this figure.

affective PEs at $t-1$ and the probability of choosing the same action at t was not statistically significant, $b = 0.027$, 95% CI [−0.026 to 0.081], $\eta^2 = 0.01$, $n = 901$, $p = .313$, and there was no evidence for an interaction by policy type, $b = -0.011$, 95% CI [−0.126 to 0.104], $\eta^2 < -0.01$, $n = 901$, $p = .847$.

H4a and H4b: Policy Escalation

As we predicted, affective PEs positively predicted the escalation of behavior (supporting H4a) but did so equally for stealing and earning (not supporting H4b). Increases in the chosen amount within the same policy from $t-1$ to t were positively and significantly associated with affective PEs at $t-1$, $b = 0.021$, 95% CI [0.010 to 0.032], $\eta^2 = 0.03$, $n = 901$, $p < .001$, but there was no evidence for an interaction by policy type, $b = -0.015$, 95% CI [−0.040 to 0.011], $\eta^2 = -0.02$, $n = 901$, $p = .261$ (see Figure 3; see Figure 1 in the online supplemental materials for descriptive visualizations including distributions of the data).

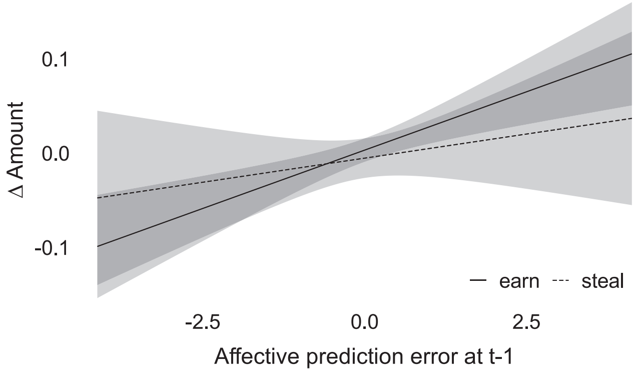
Adjusting Significant Results for Postoutcome Affect

Policy escalation showed a significant positive association with affective PEs (H4a), but we wanted to make sure that this association was best explained by feeling better than expected as opposed to merely feeling good following the outcome. Including outcome affect in the model did not change our central result—affective PEs continued to explain significant variance above and beyond outcome affect, $b = 0.029$, 95% CI [0.016 to 0.042], $\eta^2 = 0.04$, $n = 895$, $p < .001$ (see the online supplemental materials for additional analyses). Furthermore, model comparison showed that a model including postoutcome affect instead of affective PEs would be less suited to account for the data ($BF_{postOutcome} < 0.01$).

In summary, participants stole more from disliked targets, and more important, escalation of behavior (earning or stealing more points on the next trial) was greater the more a given action felt better than expected. The escalation finding survived adjustment for postoutcome affect while showing no difference between aggression and nonaggression.

Figure 3

Aggression Escalation in Experiment 1



Note. $N = 901$. Model-based predictions of difference in amount chosen for same policy from $t-1$ to t (y axis) as a function of within-participant standardized affective PEs at $t-1$ (x axis). Separate lines are fitted for each prior policy. Shaded areas reflect 95% CIs. CI = confidence interval; affective PE = affective prediction error.

Experiment 1b

This experiment mirrored Experiment 1 except for two key differences: First, participants no longer chose between earning for themselves and stealing from a target; instead, they chose between adding to and subtracting from the target's endowment without receiving any points for themselves. Subtraction thus only serves to hurt the target because it confers no benefit to the participant, making it a purer measure of aggression that also more closely aligns with the original PSAP task. Second, the target was explicitly presented as being unable to retaliate or react. Procedure, methods, and analysis remained the same, allowing us to rule out effects of personal gain and implicit fear of retaliation as directly as possible. The preregistration for this experiment can be found at https://osf.io/gj34v/?view_only=74043b2055cb4154b0b355e3f172f759.

Method

Participants

All participants were recruited in the same way and based on the same logic as in Experiment 1 (bonus not applicable). Out of 998 complete submissions, seven participants were excluded because they stated to have experienced more than zero fatal heart attacks in their lives, 36 participants were excluded because they averaged less than half (five) of the required (10) button presses. Following exclusions, this yielded a final sample of 955 participants (389 women, 538 men, 28 other, $M_{age} = 38.6$, age range = 18–77, $SD = 13.50$, 694 Democrat).

Results

Inheriting the hypotheses formulated for Experiment 1, Experiment 1b tested H1, H2, H3, and H4. Specifically, we tested whether participants would destroy more points from targets of disliked political parties (Democrats vs. Republicans; H1). We predicted that mean affective PEs would be increased for aggressive actions (H2). We further predicted that affective PEs at $t-1$ are positively associated with both choosing the same action (policy persistence; H3a) and increases in the amount chosen at t (policy escalation; H4a). In case affective PEs show no difference by policy type (i.e., no evidence for H2), we also predicted these associations to be stronger for aggressive actions (H3b and H4b, respectively).

Manipulation Check and H1: Group Bias

Again, consistent with general harm aversion, people chose to create more often than to destroy (26.7% of all choices). As predicted, however, the intergroup manipulation yielded bias in both attitudes (manipulation check) and destruction (supporting H1). People identified less with outgroup ($M_{outgroup} = 23.9$, $SD = 29.5$) compared to ingroup ($M_{ingroup} = 58.5$, $SD = 25.7$) members, $b = -34.618$, 95% CI [−36.480 to −32.755], $\eta^2 = -1.45$, $n = 955$, $p < .001$, engaging in less destruction toward targets with whose group they identified more, $b = -0.617$, 95% CI [−0.801 to −0.434], $\eta^2 = -0.19$, $n = 955$, $p < .001$.

H2: Affective PEs for Destruction Versus Creation

Contrary to our original hypothesis, but replicating Experiment 1, participants did not report significantly larger affective PEs following destruction trials compared to creation trials, $b = 0.030$, 95% CI [−0.002 to 0.061], $\eta^2 = 0.03$, $n = 931$, $p = .062$.

H3a and H3b: Policy Persistence

Contrary to our hypotheses, but replicating Experiment 1, affective PEs showed no association with subsequent policy persistence, neither across policies nor when comparing destruction to creation. The association between affective PEs at $t - 1$ and the probability of choosing the same action at t was not statistically significant, $b = -0.063$, 95% CI [−0.136 to 0.011], $\eta^2 = -0.01$, $n = 930$, $p = .093$, and there was no evidence for an interaction by policy type, $b = 0.060$, 95% CI [−0.092 to 0.213], $\eta^2 = 0.01$, $n = 930$, $p = .437$.

H4a and H4b: Policy Escalation

As we predicted, and replicating Experiment 1, affective PEs positively predicted the escalation of behavior (supporting H4a) but did so equally for destruction and creation (not supporting H4b). Increases in the chosen amount within the same policy from $t - 1$ to t were positively and significantly associated with affective PEs at $t - 1$, $b = 0.030$, 95% CI [0.011 to 0.049], $\eta^2 = 0.03$, $n = 927$, $p = .002$, but there was no evidence for an interaction by policy type, $b = -0.037$, 95% CI [−0.081 to 0.008], $\eta^2 = -0.03$, $n = 927$, $p = .104$.

Adjusting Significant Results for Postoutcome Affect

Policy escalation showed a significant positive association with affective PEs (H4a), but, as in Experiment 1, we wanted to make sure that this association was best predicted by feeling better than expected as opposed to merely feeling good following the outcome. As before, including outcome affect in the model did not change our central result—affective PEs continued to explain significant variance above and beyond postoutcome affect, $b = 0.092$, 95% CI [0.069 to 0.114], $\eta^2 = 0.08$, $n = 916$, $p < .001$ (see the online supplemental materials). Furthermore, model comparison showed that a model including postoutcome affect instead of affective PEs would be less suited to account for the data ($\text{BF}_{\text{postOutcome}} = 0.04$)

In summary, Experiment 1b replicated Experiment 1 in the absence of personal gains or justified fear of retaliation: Participants destroyed more points from disliked targets, and more important, escalation of behavior (creating or destroying more points on the next trial) was greater the more a given action felt better than expected. This finding survived adjustment for postoutcome affect while showing no difference between aggression and nonaggression.

Experiment 2

Method

Experiments 1 and 1b provided initial evidence for a predictive role of affective PEs in behavior escalation. That said, we observed a relatively low proportion of stealing or destruction (∼30%), making it a challenge to detect the predicted relationships for aggressive behavior (in isolation and relative to nonaggressive behavior). In Experiment 2, we sought to increase the proportion of aggressive behavior by making stealing more attractive, better positioning us to detect asymmetries between policies regarding their association with affective PEs. Removing the superficial equivalence of policies (i.e., equal maximum of points gained from earning and stealing) in favor of stealing was expected to increase overall stealing and change the stakes of switching policies, better positioning us to detect associations between affective PEs and stealing. Note again that there is zero outcome prediction error in all cases as participants earned exactly what they expected. The preregistration for this experiment is available at https://osf.io/t4uhv/?view_only=26093255203a459899128a1c48e8ae0b.

Participants

All participants were recruited in the same way and based on the same logic as in the previous experiments. We obtained 987 completed submissions. Nine participants were excluded because they had stated to have experienced fatal heart attacks in their lives. Thirty-six participants were excluded because they averaged less than five button presses. This resulted in a final sample of 942 participants (428 female, 496 male, 18 other, $M_{\text{age}} = 34.7$, age range = 18–81, $SD = 13.2$, 691 Democrat).

Manipulations and Procedure

This experiment was almost identical to Experiment 1, but participants could no longer choose the specific amount when choosing to earn. Instead, the amount was fixed to 1 and the option to earn was chosen by pressing "E," whereas stealing required pressing a number between 1 and 8 only. Consequently, the total number of steps to make a selection at the PSAP choice stage was reduced from two to one. From the participants' perspective, the more important difference was that the value of stealing was up to 8 times higher than the value of earning. From an experimental design perspective, we of course needed to consider whether this asymmetry in underlying outcomes could pose a threat to our inferences. We were interested in the link between affective PEs and behavior when outcome prediction errors are held constant. This implies two aspects that remain unchanged by introducing asymmetric outcomes. First, participants continued to obtain the outcome they expected, such that outcome prediction errors remained fixed to zero. Second, affective PEs would be expected to differ in terms of their mean levels depending on the size of the underlying outcomes, because increased outcomes should come with increased affective expectations and thus increased absolute deviations from affective outcomes on average. However, these changes could be independent of whether variance—within participant, over time—in affective PEs predicts subsequent behavior. Consequently, using asymmetric outcomes could still pose a challenge to inferences regarding affective PEs, if the variance in affective PEs was systematically reduced by a more constrained outcome range. Although this can only affect comparisons between earning and stealing (specifically H3b, because H4b is not testable in Studies 2 and 3), we addressed this potential confound by comparing affective PE variances for stealing and earning across study designs and found no difference between Experiments 1 and 2. Interestingly, we found the opposite-going effect comparing Experiments 1 and 3 (i.e., greater affective PE variance in Experiment 3 when there was a smaller range of outcomes; see the online supplemental materials).

Analysis

We relied on the same approach outlined in Experiment 1. The only difference was that policy escalation could only be assessed within stealing trials (and not compared across policies), because participants were not able to change the amount (i.e., escalate) when earning.

Results

Building on Experiment 1, we fixed the amount that could be gained by earning to 1, which also meant that participants could no longer increase the amount and thus escalate when choosing to earn; outcome contingencies for stealing remained unchanged. We hypothesized that we would replicate increased stealing from disliked outgroup targets (H1); that affective PEs at $t-1$ would be positively associated with policy persistence at t (H3a), but more so for stealing (H3b); and that replicating Experiment 1, affective PEs at $t-1$ would be positively associated with policy escalation at t (H4a). Because of the change in outcome structure, we focused on the predictive role of variance in affective PEs hereafter. Consequently, it was also no longer sensible to assess mean-level differences in affective PEs between earning and stealing (H2). Relatedly, differences between escalation of earning and stealing (H4b) were no longer tested either, as participants could not escalate the chosen amount in earning trials.

Manipulation Check and H1: Group Bias

The proportion of aggressive behavior (58.8% of all choices) was roughly double that observed in the previous two experiments. As in Experiment 1, participants displayed intergroup bias in both attitudes (manipulation check) and stealing (descriptively; marginally supporting H1). Participants identified less with outgroup ($M_{\text{outgroup}} = 24.3$, $SD = 29.4$) compared to ingroup ($M_{\text{ingroup}} = 59.7$, $SD = 25.1$) members, $b = -35.391$, 95% CI [−37.348 to −33.434], $\eta^2 = 1.50$, $n = 942$, $p < .001$, but only marginally stole less from targets with whom they identified more, $b = -0.206$, 95% CI [−0.421 to 0.009], $\eta^2 = -0.05$, $n = 942$, $p = .061$.

H3a and H3b: Policy Persistence

Having induced a more balanced distribution of choices, Experiment 2 supported our hypothesis that policy persistence would be linked to affective PEs (H3a), and this association was stronger in stealing versus earning (H3b). Affective PEs at $t-1$ were positively and significantly associated with the probability of choosing the same policy at t, $b = 0.129$, 95% CI [0.072 to 0.186], $\eta^2 = 0.05$, $n = 936$, $p < .001$. This was more pronounced for stealing compared to earning at $t-1$, $b = 0.143$, 95% CI [0.019 to 0.268], $\eta^2 = 0.04$, $n = 936$, $p = .024$.

H4a: Policy Escalation

Similar to Experiment 1, but restricted to stealing, affective PEs were positively associated with policy escalation (supporting H4a). Increases in the chosen amount within stealing from $t-1$ to t were positively and significantly associated with affective PEs at $t-1$, $b = 0.099$, 95% CI [0.066 to 0.132], $\eta^2 = 0.05$, $n = 720$, $p < .001$ (see Figure 4 in the online supplemental materials).

Adjusting Significant Results for Postoutcome Affect

Regarding policy persistence, neither the main effect (H3a), $b = 0.005$, 95% CI [−0.057 to 0.068], $\eta^2 = <0.01$, $n = 929$, $p = .868$, nor the interaction held when adjusting for outcome affect (H3b), $b = -0.089$, 95% CI [−0.237 to 0.060], $\eta^2 = -0.02$, $n = 929$, $p = .241$. By contrast, the corresponding main effect for outcome affect did explain significant portion of variance, exploratory ($b = 0.256$, 95% CI [0.193 to 0.320], $\eta^2 = 0.10$, $n = 928$, $p < .001$), along with a significant interaction, exploratory ($b = 0.445$, 95% CI [0.291 to 0.600], $\eta^2 = 0.12$, $n = 929$, $p < .001$) indicating that feeling good about an action increased the probability of choosing that action again relatively more for stealing compared to earning. Lastly, the result regarding policy escalation remained significant when adjusting for outcome affect (H4a), $b = 0.042$, 95% CI [0.004 to 0.081], $\eta^2 = 0.02$, $n = 716$, $p = .032$. Notably, however, models including postoutcome affect instead of affective PEs fit the data better in all cases ($\text{BFs}_{\text{postOutcome}} > 100$).

In summary, Experiment 2 replicated the findings of Experiment 1 and further supported the main effects of the hypotheses it tested. Participants marginally stole more from disliked outgroup targets, but more important, participants were more likely to persist in stealing (more so than earning) and escalated stealing when it felt better than expected. The persistence findings, however, were attenuated by adjustment for postoutcome affect, and models including postoutcome affect instead of affective PEs fit the data better.

Experiment 3

Method

Experiment 2 revealed participants stuck more with aggressive actions that felt better than expected (also relative to neutral actions) and escalated aggressive actions more when those actions felt better than expected, but these findings are correlational. Could trial-level affective signal be influenced by more incidental affect that is unrelated to the individual differences? In Experiment 3, we wanted to see whether the weighting of affective PEs during stealing is sensitive to prior experience of pleasure (schadenfreude or counter-empathy) or empathy in response to ingroup and outgroup suffering, incidental to the task. We modified the design used in Experiment 2 by introducing a (counter-)empathy versus neutral emotion induction as a between-participant manipulation. Participants were presented with vignettes about members of the matched target's group experiencing different events. Depending on condition, those events were neutral (neutral condition) or negative ([counter-]empathy condition) events, and participants provided ratings about how good these vignettes made them feel on each trial. Furthermore, we wanted to use this opportunity to replicate the previous pattern of results in a different group context. Participants were now always matched with members of groups they did not belong to themselves (i.e., there was never the option to play with an ingroup member, just relatively more or less liked outgroup members). Building on our continuous measure of target identification (in lieu of mere in- vs. outgroup), we expected that there would still be variability in how much participants identify with a given target group, allowing to conceptually replicate the previous finding in a third-party intergroup setting. The preregistration for this experiment is available at https://osf.io/wynx8/?view_only=fff0bbcefc974e4e8204542d2f22dbf6.

Participants

All participants were recruited in the same way and based on the same logic as in the previous experiments, with one important exception: We switched from U.S. political groups to groups based on nationality. Specifically, we recruited participants from across the globe (still using Prolific), as long as they identified with any nationality other than American or Chinese, such that both of those groups would constitute neutral or outgroups with varying levels of identification. We obtained 1,010 completed submissions. Three participants were excluded because they had stated to have experienced fatal heart attacks in their lives. Forty-eight participants were excluded because they averaged less than five button presses. An additional 46 participants were excluded because they had stated their nationality to be at least in part American or Asian (which may have included Chinese). This resulted in a final sample of 913 participants (536 women, 363 men, 14 other, $M_{\text{age}} = 35.05$, age range = 18–77, $SD = 13.21$, 737 British, 49 continental European nationalities, 31 African nationalities, 96 other nationalities, 461 assigned to the control condition).

Manipulations and Procedure

This experiment was similar to Experiment 2. Two notable differences were the national group identities and (counter-)empathy induction.

Demographics. Given the focus on national groups, participants were no longer asked about their identification with U.S. political parties. Instead, participants were asked about their nationality ("What is your nationality?"; response options: "American," "Chinese," and "Other"). Selecting "Other" prompted further specification in a text box. This allowed recording specific nationalities while ensuring neither American nor Chinese constituted national ingroups.

(Counter-)Empathy Vignettes. In order to induce participants with (counter-)empathy, we presented eight vignettes (neutral or negative; between-participant condition) immediately following the practice trials, and participants were told that this part of the experiment was meant to improve performance in the subsequent task. On each induction trial, participants were presented with a short vignette of an event happening to another person. The events were either (mildly) negative ("got soaked by a taxi driving through a puddle") or neutral ("tied their shoe after noticing it was untied"). The vignette included the protagonist's name, and we made sure the name was always associated with the nationality the participant was matched with (Chinese name if matched with Chinese person, e.g., "Wang-Jing"; American name if matched with American person, e.g., "Andrew"). Note that the protagonists in the stories were not presented to include the target with which the participant was paired, making this an intergroup rather than interpersonal emotion induction. We used equal proportions of male and female names and randomized the combination between names and vignettes. On the same screen, participants had to rate how good the vignette made them feel on a continuous scale from *not at all good* over *moderately good* to *extremely good* (internally coded as 0–100).

Analysis

We relied on the same approach outlined in Experiment 2. The introduction of the induction condition added two additional analyses: a manipulation check to confirm that participants felt better about the misfortunes of disliked compared to liked others, and a three-way interaction to assess the effect of the emotion induction on the relationship between affective PEs and target identification in predicting policy escalation. The model corresponding to this three-way interaction was specified as follows:

$$\begin{aligned}\Delta\text{ amount } \sim\ & b_0 + b_1\alpha + b_2\tau + (b_3 + \varepsilon_p)\zeta + \cdots \\ & + b_7\alpha^*\tau^*\zeta + \varepsilon_s + \varepsilon_t,\end{aligned} \tag{7}$$

where b represents regression weights of the following fixed effects: intercept (b_0), affective PEs (α), target identification (τ), and induction condition (ζ). Furthermore, we included weights for the two-way interactions (b_4, b_5, b_6) as statistical constituents for the three-way interaction of interest (b_7). The random effects structure included terms for repeated measurements per participant (ε_s), and trial number (ε_t).

Results

We hypothesized that we would replicate the negative association between target identification and stealing (H1), as well as the positive association between affective PEs at $t - 1$ and policy persistence (H3a)—including its asymmetry for stealing compared to earning (H3b)—as well as policy escalation at t (H4a). In addition, we predicted that the (counter-)empathy induction condition would interact with target identification and affective PEs in predicting policy escalation, such that the induction would increase the association strength between stealing escalation and affective PEs toward disliked relative to liked others (and that this group difference would be larger in the emotion relative to the neutral induction; H5).

Manipulation Check and H1: Group Bias

Similar to Experiment 2, around two thirds of choices were aggressive (67.9%). As intended, participants' (majority British) attitudes did not differ between Chinese ($M_{\text{Chinese}} = 62.8$, $SD =$ 26.4) and American ($M_{\text{American}} = 62.1$, $SD = 22.4$) targets ($b =$ 0.648, 95% CI [−0.571 to 1.866], $\eta^2 = 0.03$, $n = 913$, $p = .298$), but, in line with H1, they still stole less from targets with whom they identified more ($b = -0.230$, 95% CI [−0.423 to −0.038], $\eta^2 = -0.06$, $n = 913$, $p = .019$).

H3a and H3b: Policy Persistence

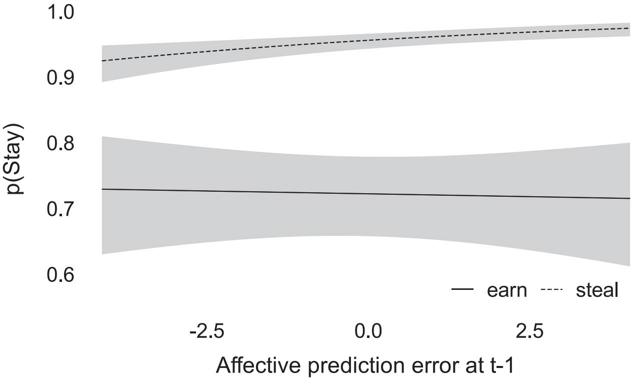
As before, and in line with our hypotheses, policy persistence correlated with affective PEs (H3a). In addition, it also showed a stronger association in stealing versus earning (H3b). Specifically, affective PEs at $t - 1$ were positively and significantly associated with the probability of choosing the same policy at t ($b = 0.085$, 95% CI [0.034 to 0.136], $\eta^2 = 0.04$, $n = 908$, $p = .001$). This was more pronounced for stealing compared to earning at $t - 1$ ($b =$ 0.162, 95% CI [0.051 to 0.274], $\eta^2 = 0.05$, $n = 908$, $p = .004$; see Figure 4).

H4a: Policy Escalation

Mirroring the previous three experiments, and restricted to stealing, affective PEs were positively associated with escalations in

Figure 4

Policy Persistence in Experiment 3



Note. $N = 908$. Model-based probability of choosing the policy chosen at $t - 1$ again at (y axis) as a function of within-participant standardized affective PEs at $t - 1$ (x axis). Separate lines are fitted for each prior policy. Affective PEs were positively associated with persistence in the same policy chosen at the previous trial, and this was even more pronounced for stealing compared to earning. However, this result was better accounted for by post-outcome affect alone (rather than by the composite prediction error). Shaded areas reflect 95% CIs. CI = confidence interval; affective PE = affective prediction error.

aggressive behavior (supporting H4a). Increases in the chosen amount within stealing from $t - 1$ to t were positively and significantly associated with affective PEs at $t - 1$ ($b = 0.096$, 95% CI [0.060 to 0.131], $\eta^2 = 0.04$, $n = 792$, $p < .001$).

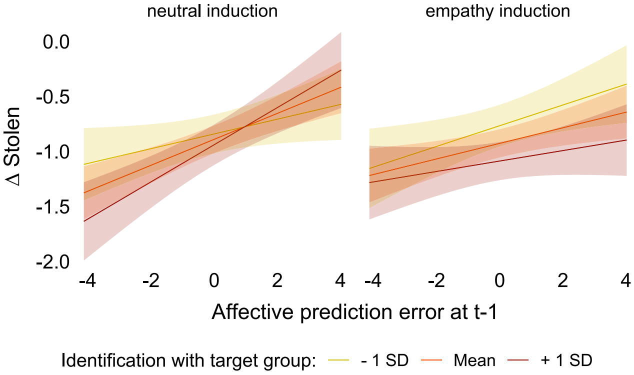
H5: Emotion Induction

The (counter-)empathy induction proved effective: people felt better about misfortunes happening to disliked others ($b = -1.732$, 95% CI [−3.163 to −0.302], $\eta^2 = 0.002$, $n = 913$, $p = .018$), and this flipped for neutral events (i.e., they felt more bad about neutral events happening to disliked others, $b = 4.368$, 95% CI [2.380 to 6.356], $\eta^2 = 0.02$, $n = 913$, $p < .001$). The key takeaway is that the more people identified with the target, the less schadenfreude/more empathy they reported.

The induction condition also affected the association between affective PEs and target identification in predicting policy escalation as hypothesized: in the (counter-)empathy induction condition (compared to the neutral condition), affective PEs predicted steeper increases in stealing from disliked than liked targets ($b_{\text{interaction}} = -0.073$, 95% CI [−0.143 to −0.003], $\eta^2 = -0.03$, $n = 792$, $p = .040$). Note, however, that the simple slope of the counter-empathy condition was not significant, $b_{\text{simpleslope}} = -0.025$, 95% CI [−0.074 to 0.025], $\eta^2 = -0.01$, $n = 390$, $p = .333$ (see Figure 5; see Figure 7 in the online supplemental materials for underlying distribution). In other words, while the association between stealing and affective PEs did not depend on target identification in general (replicating the previous three experiments; see the online supplemental materials), participants induced with counter-empathy (compared to the neutral induction) were relatively more likely to escalate stealing that felt better than expected specifically when it was directed at disliked others relative to others about whom they felt more positively inclined. The relationship between affective PEs and policy escalation did not vary as much by target dislike in the neutral condition. That said, when examining slopes across the neutral and counter-empathy induction conditions, it is apparent that rather than counter-empathy acting as an accelerant of the affective PE-by-stealing association for disliked targets, the pattern is better characterized as empathy acting as an inhibitor of the affective PE-by-stealing association for liked targets (i.e., a flattening of the relationship between target feelings and escalation for the liked targets). Moreover, much of the variability in this sample spans different levels of reductions in the amount stolen from one trial to the next—due to the unequal distribution of choices (see Figure 7 in the online supplemental materials for the complete range of changes in the amount chosen). In those cases, a steeper slope for disliked targets corresponds to a steeper decline in deescalation and thus relatively more escalatory behavior toward disliked targets.

Figure 5

Effects of Empathy Induction in Experiment 3



Note. $N = 792$. Model-based difference in amount stolen from $t - 1$ to t (y axis) as a function of within-participant standardized affective PEs at $t - 1$ (x axis), target identification (yellow = *−1 SD*, orange = *M*, purple = *+1 SD*), and emotion induction condition (control condition, left panel; empathy induction, right panel). Following the empathy induction, the association between stealing feeling better than expected and doing more of it differentiated by group identification: stealing that felt better than expected was escalated less when directed at members of liked groups. Shaded areas reflect 95% CIs. CI = confidence interval; affective PE = affective prediction error.

Adjusting Significant Results for Postoutcome Affect

The main effect regarding policy persistence was no longer significant when adjusting for postoutcome affect (H3a), $b = 0.003$, 95% CI [−0.055 to 0.061], $\eta^2 < 0.01$, $n = 908$, $p = .917$, and the interaction by policy type did not hold either (H3b), $b = -0.051$, 95% CI [−0.182 to 0.081], $\eta^2 = -0.02$, $n = 907$, $p = .450$. The corresponding interaction with postoutcome affect again explained a significant portion of variance (exploratory, $b = 0.458$, 95% CI [0.317 to 0.599], $\eta^2 = 0.15$, $n = 907$, $p < .001$), suggesting that feeling good about an action increased the probability of choosing that action again relatively more for stealing compared to earning. The result regarding policy escalation was no longer significant when adjusting for postoutcome affect (H4a), $b = 0.032$, 95% CI [−0.009 to 0.072], $\eta^2 = 0.01$, $n = 791$, $p = .128$. Lastly, the central finding of the intervention effect remained unchanged when adjusting for postoutcome affect (H5), $b_{\text{interaction}} = -0.086$, 95% CI [$-0.167$ to -0.005], $\eta^2 = -0.04$, $n = 791$, $p = .038$; $b_{\text{simpleslope}} = -0.029$, 95% CI [$-0.085$ to -0.028], $\eta^2 = -0.01$, $n = 389$, $p = .318$, following the emotion induction, stealing that felt better than expected was still escalated less when directed at members of liked groups. Notably, in all cases, models solely including postoutcome affect instead of affective PEs provided a better fit to the data ($\text{BFs}_{\text{postOutcome}} > 100$).

Experiment 3b

Similar to Experiment 1b in relation to Experiment 1, this experiment mirrored Experiment 3 except for two key differences: First, participants no longer chose between earning for themselves and stealing from another player; instead, they chose the amount to subtract from another player's points (i.e., destroy) without receiving any points for themselves (one policy in total). Second, the other player was explicitly presented as being unable to retaliate or react. Procedure, methods, and analysis remained the same, allowing us to rule out effects of personal gain and fear of retaliation as directly as possible. The preregistration for this experiment is available at https://osf.io/yc2p8/?view_only=db9989c6210e4ef29ea7983d97609d52.

Method

Participants

All participants were recruited in the same way and based on the same logic as in Experiment 3 (bonus not applicable), with one important exception: We now only recruited participants from a single country (United States), such that American targets would constitute an ingroup. We obtained 1,000 completed submissions. Sixteen participants were excluded because they had stated to have experienced fatal heart attacks in their lives. Ninety-five participants were excluded because they averaged less than five button presses. An additional two participants were excluded because they had stated a nationality other than U.S. American. This resulted in a final sample of 887 participants (349 women, 523 men, 15 other, $M_{\text{age}} = 38.86$, age range = 18–85, $SD = 13.70$, 462 assigned to the control condition).

Results

We hypothesized that we would replicate the negative association between target identification and aggression (H1), as well as the positive association between affective PEs at $t - 1$ and policy escalation at t (H4a). In addition, we predicted that the (counter-)empathy induction condition would interact with target identification and affective PEs in predicting policy escalation, such that the induction would increase the association strength between destruction escalation and affective PEs toward disliked relative to liked others (and that this group difference would be larger in the emotion relative to the neutral induction; H5).

Manipulation Check and H1: Group Bias

Participants displayed intergroup bias in attitudes (manipulation check) but not in aggressive behavior (not supporting H1). Participants identified less with Chinese ($M_{\text{Chinese}} = 65.0$, $SD = 27.0$) compared with American ($M_{\text{American}} = 71.4$, $SD = 22.7$) nationals ($b = -6.434$, 95% CI [-7.878 to -4.991], $\eta^2 = 2.81$, $n = 887$, $p < .001$), but showed no evidence for a difference in aggression depending on how much they identified with the target player ($b = -0.115$, 95% CI [-0.311 to 0.081], $\eta^2 = -0.04$, $n = 887$, $p = .249$).

H4a: Policy Escalation

Unlike Experiment 3, affective PEs showed no association with escalation (not supporting H4a). Increases in the chosen destruction amount from $t - 1$ to t were not significantly linked to affective PEs at $t - 1$ ($b = 0.012$, 95% CI [-0.011 to 0.035], $\eta^2 = 0.01$, $n = 868$, $p = .303$).

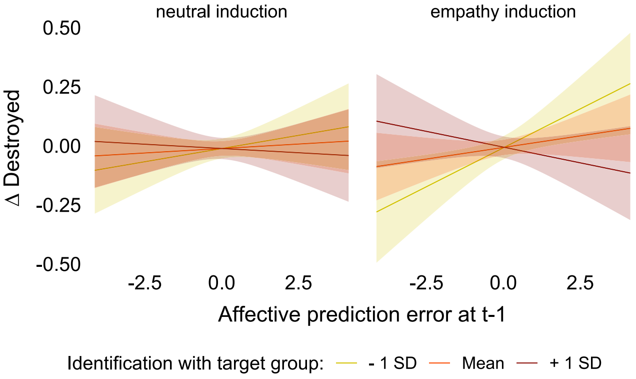
H5: Emotion Induction

The (counter-)empathy induction proved marginally effective: people felt better about misfortunes happening to disliked others ($b = -1.495$, 95% CI [-3.179 to 0.189], $\eta^2 = 0.01$, $n = 887$, $p = .082$), and this flipped for neutral events ($b = 6.061$, 95% CI [3.802 to 8.320], $\eta^2 = 0.03$, $n = 887$, $p < .001$). Again, the more people identified with the target, the less schadenfreude/more empathy they reported.

The induction condition affected the association between affective PEs and target identification in predicting policy escalation partially as hypothesized: in the (counter-)empathy induction condition affective PEs predicted steeper increases in destruction toward disliked targets, $b_{\text{simpleslope}} = -0.046$, 95% CI [$-0.082$ to -0.011], $\eta^2 = -0.03$, $n = 414$, $p = .012$ (see Figure 6; see Figure 9 in the online supplemental materials for underlying distribution) but not significantly more than in the neural condition ($b_{\text{interaction}} = -0.032$, 95% CI [$-0.079$ to -0.016], $\eta^2 = -0.02$, $n = 868$, $p = .188$). In other words, participants induced with counter-empathy were

Figure 6

Effects of Empathy Induction in Experiment 3b



Note. $N = 868$. Model-based difference in amount destroyed from $t - 1$ to t (y axis) as a function of within-participant standardized affective PEs at $t - 1$ (x axis), target identification (yellow = *−1 SD*, orange = *M*, purple = *+1 SD*) and emotion induction condition (control condition, left panel; empathy induction, right panel). Following the empathy induction, the association between destruction feeling better than expected and doing more of it differentiated by group identification: destruction that felt better than expected was escalated less when directed at members of liked groups. Shaded areas reflect 95% CIs. CI = confidence interval; affective PE = affective prediction error.

more likely to escalate destruction that felt better than expected specifically when it was directed at disliked others relative to others about whom they felt more positively inclined.

Adjusting Significant Results for Postoutcome Affect

The main effect regarding policy escalation turned significant when adjusting for postoutcome affect (in the opposite direction of H4a), $b = -0.045$, 95% CI [-0.072 to -0.018], $\eta^2 = -0.03$, $n = 865$, $p = .001$. As for the intervention effect, the interaction term turned significant (in the hypothesized direction) when adjusting for postoutcome affect, and the simple slope remained significant (H5), $b_{\text{interaction}} = -0.065$, 95% CI [$-0.119$ to -0.010], $\eta^2 = -0.04$, $n = 868$, $p = .020$; $b_{\text{simpleslope}} = -0.088$, 95% CI [$-0.129$ to -0.047], $\eta^2 = -0.06$, $n = 412$, $p < .001$, following the emotion induction, destruction that felt better than expected was still escalated less when directed at members of liked groups. In all cases, a model including postoutcome affect provided a better fit to the data compared to a model including affective PEs instead ($\text{BFs}_{\text{postOutcome}} > 10$).

General Discussion

We investigated the role of affective PEs in how people may overcome their aversion to aggression to harm (disliked) others. Across five preregistered experiments and over 4,600 participants, we found that affective PEs at $t - 1$ predicted escalation of the same behavior, both aggressive and nonaggressive, at t (Experiments 1, 1b, 2, and 3). Affective PEs at $t - 1$ were also positively associated with persisting in both behaviors at t (Experiments 2 and 3), and this association was sometimes stronger for aggressive versus nonaggressive behaviors (Experiments 2 and 3). In other words, people were even more likely to take an action that felt better than expected if that action was aggressive. Building on these associations, we looked at how presenting participants with others' misfortunes—a process that reliably makes many people feel relatively good (bad) about the misfortunes of disliked (liked) targets—changes the aforementioned relationship between affective PEs and aggression. Specifically, we predicted that experiencing schadenfreude for the disliked group would increase the relationship between affective PEs and aggression escalation more for disliked than liked targets (and relative to the neutral induction condition). While this hypothesis was technically supported (interaction in Experiment 3; simple slope in Experiment 3b) the results are better characterized as follows: after experiencing empathy for liked groups people were relatively less likely to escalate aggression that felt better than expected if it was directed specifically at liked relative to disliked targets. In other words, experiencing empathy for the target's group appears to have "put the brakes on" affective PEs' prediction of aggression escalation.

Affective PEs are not merely an affective mirror image of prediction errors as they are usually investigated (see Vollberg & Sander, 2024). While we have provided a potentially generative framework to study behavior in domains like aggression, it is important to reemphasize the differences between affective and outcome prediction errors. When making predictions about a lottery payout, both the prediction and the actual payout refer to outcomes in the environment. That is, a payout of five dollars, for example, is subject to consensus. Emotion does not have the same ground truth around which to calibrate; five dollars feels differently to us when we are in the bottom versus top decile of the income distribution. Consequently, the structure with which outcome prediction errors have been shown to inform beliefs over time (e.g., to learn the true win probability of a lottery) should not be expected in affective PEs. This does not mean that affective PEs contain no signal; the signal we observe may just be more temporally transient and noisy (i.e., less dependence across trials). With this difference in mind, we have highlighted an important tension between affective forecasting and reward learning throughout this article. Specifically, from a reward learning perspective, we might approach the studies presented here expecting affective PEs to matter—after all, prediction errors are at the core of reward learning. From an affective forecasting perspective, however, we might not be all that interested in whether somebody felt better than expected if what they do next is sufficiently predicted by how good they felt following an outcome, independent of their prior expectation. In light of this tension, the results presented here are mixed: feeling better/worse than expected sometimes explains variance above and beyond merely feeling good/bad (i.e., for several aggression escalation findings) and sometimes it does not (i.e., for most aggression persistence findings). Rather than constituting a threat, we believe this to be generative for the nascent research program on predictive roles of affective PEs. That is because our findings provide both a proof-of-concept in the cases where affective PEs do predict behavior above and beyond postoutcome affect, while also urging caution in using composites without explicitly testing more parsimonious models based on their constituent parts first.

Limitations

There are several limitations to the present investigation. Participants in our experiments could display aggressive behaviors by stealing or destroying virtual resources from an ostensible target player. Although those virtual resources were reflected in participants' bonus payments (Experiments 1–3), the resources were thus not actually stolen or destroyed from another player. We also did not test other real-world forms of aggression such as physical harm. Furthermore, the "interaction" with the target was almost completely anonymous. Our data cannot speak directly to how people would behave in less anonymous settings or with higher stakes on the form of aggression and retaliation (e.g., administering electric shocks). Reassuringly, however, participants' predicted affect ratings suggest even this low-stakes operationalization of aggression to have been relatively aversive (see Figure 2A; see Figures 2A, 3A, and 4A in the online supplemental materials).

Another limitation concerns the temporal structure we chose for the experimental task. Participants provided their affective predictions after deciding on a certain action along with the corresponding amount. Although participants still had to complete a task to actually obtain the chosen outcome, they had committed to taking the chosen action prior to making their affective predictions. Reassuringly, the presence of affective PEs—which decreased over time (see Figure 1C; see Figures 2C, 3C, 5C, and 8C in the online supplemental materials)—suggests that this postdecision state still differed from the postoutcome state. In other words, at least affectively, the commitment appears to have been partial rather than complete; otherwise, affective responses would have been equal across postdecision and postoutcome states, such that their difference would be near zero across the task. Nonetheless, it remains an open question how closely the postdecision state resembles the predecision state for the purposes of the task.

Psychologically, we might wonder whether affective PEs reflect genuinely reported emotions as opposed to motivated responses. Participants might just say that they felt better than expected to make subsequent escalation look more rational (i.e., post hoc rationalization). Although we cannot identify why participants chose what they chose, the association with affective PEs is unlikely to stem from post hoc rationalization alone, because participants in our emotion induction conditions sometimes reported feeling better than expected about aggression toward liked others that was not subsequently escalated.

We present evidence consistent with the role of affective PEs as a learning signal independent of outcome prediction errors. In the candidate context of aggression, affective PEs may constitute an algorithmic tool to solve the computational problem of aggressing only under exceptional circumstances. Based on the assumption that intergroup competition poses one such exception, we further provide evidence that this affective learning signal is sensitive to manipulation of intergroup emotion (i.e., [counter-]empathy). Throughout our analyses, we invite caution in using affective PEs to predict behavior as postoutcome affect may sometimes be more (parsimoniously) predictive. More generally, the research presented here provides a novel framework with which to study learning beyond outcome prediction errors that are held constant or set to zero.

Constraints on Generality

Our experiments captured intergroup dynamics using target identification ratings about political (U.S. context; Experiments 1, 1b, and 2) and national (United States and China; Experiments 3 and 3b) group members, recruiting participants from across the globe (Experiment 3). Although this is a step toward transcending limited target populations and group-specific idiosyncrasies, we still relied on convenience sampling, and there are innumerable group contexts beyond the ones studied here. Consequently, we cannot know if and how exactly our pattern of results translates to groups and individuals beyond those sampled here.

References

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Bandura, A. (1976). Social learning analysis of aggression. In E. Ribes Inesta & A. Bandura (Eds.), *Analysis of delinquency and aggression* (pp. 203–231). Erlbaum.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Journal of Neuroscience*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), Article 9. https://doi.org/10.5334/joc.10

Chang, L. W., Krosch, A. R., & Cikara, M. (2016). Effects of intergroup threat on mind, brain, and behavior. *Current Opinion in Psychology*, *11*, 69–73. https://doi.org/10.1016/j.copsyc.2016.06.004

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of affective decision making: How do feelings predict choice? *Psychological Science*, *27*(6), 763–775. https://doi.org/10.1177/0956797616634654

Cherek, D. R., Moeller, F. G., Schnapp, W., & Dougherty, D. M. (1997). Studies of violent and nonviolent male parolees: I. Laboratory and psychometric measurements of aggression. *Biological Psychiatry*, *41*(5), 514–522. https://doi.org/10.1016/S0006-3223(96)00059-5

Chester, D. S. (2017). The role of positive affect in aggression. *Current Directions in Psychological Science*, *26*(4), 366–370. https://doi.org/10.1177/0963721417700457

Chester, D. S., & DeWall, C. N. (2016). The pleasure of revenge: Retaliatory aggression arises from a neural imbalance toward reward. *Social Cognitive and Affective Neuroscience*, *11*(7), 1173–1182. https://doi.org/10.1093/scan/nsv082

Cikara, M. (2015). Intergroup schadenfreude: Motivating participation in collective violence. *Current Opinion in Behavioral Sciences*, *3*, 12–17. https://doi.org/10.1016/j.cobeha.2014.12.007

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, *22*(3), 306–313. https://doi.org/10.1177/0956797610397667

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125. https://doi.org/10.1016/j.jesp.2014.06.007

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations. *Perspectives on Psychological Science*, *9*(3), 245–274. https://doi.org/10.1177/1745691614527464

Collins, A. G., & Shenhav, A. (2021). Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, *47*(1), 104–118. https://doi.org/10.1038/s41386-021-01126-y

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325. https://doi.org/10.1073/pnas.1408988111

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7. https://doi.org/10.1037/a0025071

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4, Pt. 1), 377–383. https://doi.org/10.1037/h0025589

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Engelmann, J. B., Meyer, F., Ruff, C. C., & Fehr, E. (2019). The neural circuitry of affect-induced distortions of trust. *Science Advances*, *5*(3), Article eaau3413. https://doi.org/10.1126/sciadv.aau3413

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, *19*(12), 1727–1732. https://doi.org/10.1038/nn.4426

Golub, S. A., Gilbert, D. T., & Wilson, T. D. (2009). Anticipating one's troubles: The costs and benefits of negative expectations. *Emotion*, *9*(2), 277–281. https://doi.org/10.1037/a0014716

Hartig, F. (2018). *Package DHARMa* (Version 0.3.2.0) [Computer software]. https://CRAN.R-project.org/package=DHARMa

Heffner, J., Son, J.-Y., & FeldmanHall, O. (2021). Emotion prediction errors guide socially adaptive behavior. *Nature Human Behavior*, *5*(10), 1391–1401. https://doi.org/10.1038/s41562-021-01213-6

Hogg, M. A. (1993). Group cohesiveness: A critical review and some new directions. *European Review of Social Psychology*, *4*(1), 85–111. https://doi.org/10.1080/14792779343000031

Juechems, K., & Summerfield, C. (2019). Where does value come from? *Trends in Cognitive Sciences*, *23*(10), 836–850. https://doi.org/10.1016/j.tics.2019.07.012

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: Neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3771–3786. https://doi.org/10.1098/rstb.2008.0155

Levy, D., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Mellers, B., & Mcgraw, A. P. (2001). Anticipated emotions as guides to choice. *Current Directions in Psychological Science*, *10*(6), 210–214. https://doi.org/10.1111/1467-8721.00151

Mellers, B., Schwartz, A., & Ritov, D. (1999). Emotion-based choice emotion-based choice. *Journal of Experimental Psychology: General*, *128*(3), 332–345. https://doi.org/10.1037/0096-3445.128.3.332

Meyvis, T., Ratner, R. K., & Levav, J. (2010). Why don't we learn to accurately forecast feelings? How misremembering our predictions blinds us to past forecasting errors. *Journal of Experimental Psychology: General*, *139*(4), 579–589. https://doi.org/10.1037/a0020285

Mitsuko, W.-U., Neir, E., & Uchida, N. (2017). Neural circuitry of reward prediction error. *Annual Review of Neuroscience*, *40*(1), 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Nielsen, L., Knutson, B., & Carstensen, L. L. (2008). Affect dynamics, affective forecasting, and aging. *Emotion*, *8*(3), 318–330. https://doi.org/10.1037/1528-3542.8.3.318

Phelps, E. A., Lempert, K. M., & Sokol-Hessner, P. (2014). Emotion and decision making: Multiple modulatory neural circuits. *Annual Review of Neuroscience*, *37*(1), 263–287. https://doi.org/10.1146/annurev-neuro-071013-014119

Pinker, S. A. (2011). *The Better Angels of our nature: The decline of violence in history and its causes*. Penguin.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Sander, D., & Nummenmaa, L. (2021). Reward and emotion: An affective neuroscience approach. *Current Opinion in Behavioral Sciences*, *39*, 161–167. https://doi.org/10.1016/j.cobeha.2021.03.016

Schultz, W., Dayan, P., & Montague, R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.

Van Erp, A. M., & Miczek, K. A. (2000). Aggressive behavior, increased accumbal dopamine, and decreased cortical serotonin in rats. *The Journal of Neuroscience*, *20*(24), 9320–9325. https://doi.org/10.1523/JNEUROSCI.20-24-09320.2000

Vollberg, M. C., & Sander, D. (2024). Hidden reward: Affect and its prediction errors as a window into subjective value. *Current Directions in Psychological Science*, *14*(3), 131–134. https://doi.org/10.1177/09637214231217678

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, *14*(3), 131–134. https://doi.org/10.1111/j.0963-7214.2005.00355.x

Woolf, L. M., & Hulsizer, M. R. (2004). Hate groups for dummies: How to build a successful hate-group. *Humanity & Society*, *28*(1), 40–62. https://doi.org/10.1177/016059760402800105

Zhong, C.-B., Strejcek, B., & Sivanathan, N. (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology*, *46*(5), 859–862. https://doi.org/10.1016/j.jesp.2010.04.003

Zimbardo, P. G. (1995). The psychology of evil: A situationist perspective on recruiting good people to engage in anti-social acts. *Japanese Journal of Social Psychology*, *11*(2), 125–133. https://doi.org/10.14966/jssp.KJ00003724682

Received August 7, 2023
Revision received January 29, 2024
Accepted February 11, 2024 ■