SUPPLEMENTAL MATERIALS

AFFECTIVE PREDICTION ERRORS IN PERSISTENCE AND ESCALATION OF AGGRESSION

Overview of Studies

Experiment: / *Design feature:*	Experiment 1	Experiment 1b	Experiment 2	Experiment 3	Experiment 3b
Target group	Democrat / Republican	Democrat / Republican	Democrat / Republican	American / Chinese	American / Chinese
Own group	Democrat / Republican	Democrat / Republican	Democrat / Republican	Any except target (primarily British)	US Americans
Options / policy	1-8 for both earning and stealing	1-8 for both creating and destroying	1 for earning, 1-8 for stealing	1 for earning, 1-8 for stealing	1-8 for destroying
Manipulation beyond group	None	None	None	(Counter-)empathy vs. neutral induction	(Counter-)empathy vs. neutral induction
Passive target	No	Yes	No	No	Yes

Supplemental Table 1.
Overview of experimental design across Experiments 1-3. All three experiments used the same task structure but differed with respect to the groups, outcome structure, or additional manipulations involved. Experiments 1, 1b, and 2 asked self-identified Democrats or Republicans to make choices affecting a Democrat or Republican (ostensible targets randomly allocated to group identity), but the outcome structure differed between the experiments. In Experiment 1, participants could gain the same number of points whether they chose to earn or to steal. In Experiment 1b, earning and stealing was replaced with creating and destroying in order to replicate the finding of Experiment 1 in absence of personal gain. Furthermore, the target was passive and could not retaliate, thereby ruling out justified concerns about potential retaliation. Although aggressive behavior occurred around 30% of the time in Experiments 1 and 1b, Experiment 2 increased the prevalence of aggression by restricting the amount when choosing to earn to just 1 point. Experiment 3 inherited this outcome structure but differed in two other ways: participants were now led to believe to be interacting with Chinese or American targets, neither of which would correspond to their own national identity; furthermore, participants were randomly allocated to be exposed to neutral or (counter-)empathy-inducing events happening to members of the target group. Similar to Experiment 1b relative to Experiment 1, Experiment 3b replicated Experiment 3 in absence of personal gain and fear of retaliation (while also switching the population to US Americans).

Supplemental analyses

Evolution of affective prediction errors over time

The present investigation focuses only on affective prediction errors (affective PEs) immediately preceding a trial and uses them to predict choices on a given trial. This contrasts with other reward learning approaches that focus on how people incrementally learn what outcomes in the environment to expect over time. Emotions are thought to reflect the subjective appraisal of those outcomes, precluding equivalent calibration around a ground truth over time. In light of this difference, we focused on affective PEs as trial-to-trial predictors of behavior, putting less emphasis on the coherence of affective PEs across all trials. Nonetheless it is possible to explore the temporal evolution of affective PEs, too. The trajectory of affective PEs over time can be analyzed in numerous ways, and we sought to address two questions: (1) do people get better at predicting their affective responses over time, and (2) do learning rates differ as a function of key variables in the experiment (i.e., policy type or group identification)?

Affective prediction errors over time

Experiment 1

Average signed affective PEs across participants did not change across trials in Experiment 1 (b = -0.005; 95% CI, -0.027 to 0.177; η^2 > -0.01; p = 0.686). When looking at unsigned affective PEs across participants, however, there was significant reduction over time (b = -0.133; 95% CI, -0.151 to -0.116; η^2 = -0.02; p < 0.001).

Experiment 1b

Average signed affective PEs across participants did not change across trials in Experiment 1b (b = 0.014; 95% CI, -0.010 to 0.037; η^2 > 0.01; p = 0.251). When looking at unsigned affective

PEs across participants, however, there was again a significant reduction over time (b = -0.069; 95% CI, -0.088 to -0.051; η^2 = -0.01; $p < 0.001$).

Experiment 2

Average signed affective PEs increased across trials in Experiment 2 (b = 0.039; 95% CI, 0.015 to 0.064; η^2= 0.004; $p = 0.002$). Unsigned affective PEs decreased over time (b = -0.109; 95% CI, -0.128 to -0.089; η^2 = -0.01; $p < 0.001$).

Experiment 3

Average signed affective PEs showed no significant increase across trials in Experiment 3 (b = -0.026; 95% CI, -0.053 to 0.001; η^2 > -0.01; $p = 0.055$). Unsigned affective PEs did show a reduction over time (b = -0.157; 95% CI, -0.177 to -0.136; η^2 = -0.02; $p < 0.001$).

Experiment 3b

Average signed affective PEs showed no significant increase across trials in Experiment 3b (b = -0.011; 95% CI, -0.035 to 0.013; η^2 > -0.01; $p = 0.356$). Unsigned affective PEs did show a reduction over time (b = -0.058; 95% CI, -0.077 to -0.039; η^2 = -0.01; $p < 0.001$).

Summary

Taken together, it appears that, on average, signed affective PEs did not change in magnitude across time (i.e., no relative increase in under- or underestimations over time; though see Experiment 2). In terms of unsigned affective PEs, however, there is a tendency for people to report decreasing affective PEs over time. In combination, this suggests that while people maintained a steady ratio of underestimations to underestimations across time at the sample level, they also tended to get better at predicting their affective responses.

Learning rates by policy and group identification

For the present purposes, learning rate refers to the fraction of a prediction error at t relative to the size of the prediction error at t-1. The learning rate on a given trial was defined as:

$$LR_t = \frac{|aPE_{t-1}| - |aPE_t|}{|aPE_{t-1}|}$$

Note that a learning rate is most interpretable within a given policy: although a participant's affective predictions about aggression may be informed by affective prediction errors about non-aggression and vice versa, we looked at the impact of affective prediction errors on subsequent predictions (learning rate) for choices of the same policy type (i.e., the effect of a prediction error on the next prediction error about the same policy type).

Experiment 1

Learning rates showed no difference by target identification in Experiment 1 (b = -0.080; 95% CI, -0.272 to 0.113; η^2 = -0.01; p = 0.417) and there was no interaction by policy type (b = -0.004; 95% CI, -0.308 to 0.300; η^2 > -0.01; p = 0.979).

Experiment 1b

Learning rates showed no difference by target identification in Experiment 1b (b = -0.005; 95% CI, -0.197 to 0.186; η^2 > -0.01; p = 0.958) and there was no interaction by policy type (b = -0.234; 95% CI, -0.590 to 0.121; η^2 = -0.04; p = 0.197).

Experiment 2

Learning rates showed no difference by target identification in Experiment 2 (b = -0.015; 95% CI, -0.293 to 0.264; η^2 > -0.01; p = 0.919) and there was no interaction by policy type (b = -0.228; 95% CI, -0.553 to 0.097; η^2 = -0.04; p = 0.170).

Experiment 3

As in previous experiments, learning rates showed no difference by target identification in Experiment 3 (b = -0.123; 95% CI, -0.428 to 0.182; η^2 = -0.02; p = 0.429) and there was no interaction by policy type (b = -0.220; 95% CI, -0.542 to 0.102; η^2 = -0.04; p = 0.181).

Experiment 3b

Lastly, in Experiment 3b, learning rates showed no difference by target identification (b = -0.104; 95% CI, -0.308 to 0.099; η^2 = -0.02; p = 0.316) and there was only one policy type (destruction).

Summary

Based on these analyses, learning rates did not differ depending on how much participants identified with the group of the target and there was no evidence for an interaction with the chosen policy (non-aggressive or aggressive) either.

Variance of affective prediction errors by outcome structure

One concern we had about changing the outcome structure between experiments (such that participants could only ever choose one point for the non-aggressive action) was that this change would truncate variance in affective PEs on non-aggressive trials compared to aggressive trials in Experiments 2 and 3 relative to Experiment 1. The ratio of affective PE variances for earning versus stealing trials showed no significant difference when comparing Experiment 1 to Experiment 2 (b = -20.360; 95% CI, -54.621 to 13.901; η^2 = -0.09; p = 0.244). Interestingly, affective PE variance for stealing compared to earning was significantly *smaller* in Experiment 3 than in Experiment 1 (b = -35.854; 95% CI, -69.742 to -1.966; η^2 = -0.15; p = 0.038). Thus, while the results for differences in affective PE variance by outcome structure are somewhat inconsistent,

we find the opposite of what we were concerned about: affective PE variance was reduced when participants had relatively *larger* outcome ranges.

Affective prediction errors and target identification

Experiment 1

General levels of affective prediction errors did not depend on how much participants identified with the group of the target (b = 0.0005; 95% CI, -0.014 to 0.014; $\eta^2 < 0.01$; n = 901; $p = 0.950$). Policy persistence as predicted by affective prediction errors did not interact with target identification (b = -0.044; 95% CI, -0.114 to 0.026; $\eta^2 = -0.02$; n = 901; $p = 0.221$) and there was no evidence for an interaction by policy type (b = 0.030; 95% CI, -0.076 to 0.136; $\eta^2 = 0.01$; n = 901; $p = 0.574$). The relationship between policy escalation and affective prediction errors did not interact with target group identification either (b = -0.002; 95% CI, -0.013 to 0.010; $\eta^2 > -0.01$; n = 901; $p = 0.787$).

Experiment 1b

General levels of affective prediction errors did not depend on how much participants identified with the group of the target (b = 0.003; 95% CI, -0.012 to 0.017; $\eta^2 < 0.01$; n = 931; $p = 0.727$). Policy persistence as predicted by affective prediction errors negatively interacted with target identification within non-aggressive trials (b = -0.138; 95% CI, -0.232 to -0.044; $\eta^2 = -0.03$; n = 930; $p = 0.004$) showing a significant positive relationship in aggressive trials (b = 0.164; 95% CI, 0.022 to 0.306; $\eta^2 = 0.03$; n = 901; $p = 0.024$). Taken at face value, this would counterintuitively suggest that participants in this experiment stuck more to aggressive behavior that felt better than expected and less to non-aggressive behavior when directed at members of *liked* groups. The relationship between policy escalation and affective prediction errors did not

interact with target group identification (b = -0.002; 95% CI, -0.021 to 0.017; η^2 > -0.01; n = 927; p = 0.857).

Experiment 2

General levels of affective prediction errors did not depend on how much participants identified with the group of the target (b = 0.0004; 95% CI, -0.013 to 0.014; η^2 < 0.01; n = 935; p = 0.957). Policy persistence as predicted by affective prediction errors did not interact with target identification (b = 0.009; 95% CI, -0.074 to 0.091; η^2 = < 0.01; n = 935; p = 0.834) and there was no evidence for an interaction by policy type (b = -0.092; 95% CI, -0.205 to 0.022; η^2 = -0.04; n = 935; p = 0.115). The relationship between policy escalation and affective prediction errors did not interact with target group identification (b = 0.013; 95% CI, -0.021 to 0.046; η^2 < 0.01; n = 719; p = 0.459).

Experiment 3

General levels of affective prediction errors did not depend on how much participants identified with the group of the target (b = 0.0007; 95% CI, -0.013 to 0.015; η^2 < 0.01; n = 908; p = 0.922). Policy persistence as predicted by affective prediction errors did not interact with target identification (b = 0.024; 95% CI, -0.054 to 0.102; η^2 = 0.01; n = 908; p = 0.550) and there was no evidence for an interaction by policy type (b = -0.014; 95% CI, -0.118 to 0.091; η^2 > -0.01; n = 908; p = 0.797). The relationship between policy escalation and affective prediction errors did not interact with target group identification (b = 0.010; 95% CI, -0.025 to 0.045; η^2 < 0.01; n = 792; p = 0.561).

Experiment 3b

General levels of affective prediction errors did not depend on how much participants identified with the group of the target (b < 0.0001; 95% CI, -0.014 to 0.014; η^2 < 0.01; n = 870; p

= 0.999). Policy persistence does not apply because there was only one policy to choose from in Experiment 3b. Policy escalation, however, showed an interaction with target identification (b = -0.028; 95% CI, -0.051 to -0.005; η^2 = -0.02; n = 868; p = 0.018). In other words, participants in this experiment escalated destruction that felt better than expected less towards members of more liked groups, which is in line with the simple slope in the experimental condition.

Summary

While the analyses reported in the manuscript reveal general associations of affective PEs with behavior, as well as disproportionate aggression against disliked others, they do not contain target identification as a term interacting with affective prediction errors themselves. The additional analyses documented here show no evidence for a consistent association between affective PEs and group identification beyond the patterns related to the emotion induction condition in Experiments 3 and 3b.

Robustness checks

To ensure that the results reported in the manuscript—especially regarding the core findings—are robust and not due to specific analysis choices, we conducted three additional analyses. All analyses were based on the models reported in the manuscript but added predictors or changed the transformation of existing ones. Added covariates included age, gender, and group membership as applicable (i.e., political party in Experiments 1, 1b & 2, nationality in Experiments 3 and 3b). In addition to these alternative specifications of pre-registered models, we also provide separate analyses that probe the role of the constituents of affective prediction errors. As noted in the manuscript, there are at least two approaches to doing this: (1) joint regression models including the prediction error along with one of its constituent parts, and (2) model comparison pitting a model including each constituent against a model including the prediction error. Given the potential for suppression in the former approach due to the interrelatedness of prediction errors and their constituent parts, the latter approach (model comparison) may be considered more informative. See Supplemental Table 2 for details.

Model: / *Effect:*	*Linear mixed model (LMM), affective PEs standardized within;* ***as reported***	*LMM as reported + covariate adjusted (age, gender, group)*	*LMM as reported except predictors standardized across participants*	*LMM as reported except predictors as unstandardized values*	*LMM as reported adjusted for post-outcome affect (OA)*	*LMM as reported adjusted for predicted affect (PA)*
H1: aggression ~ target ID						
Experiment 1	-0.437 to -0.071	-0.424 to -0.059	NA	-0.015 to -0.002	NA	NA
Experiment 1b	-0.801 to -0.434	-0.799 to -0.430	NA	-0.027 to -0.015	NA	NA
Experiment 2	-0.421 to 0.009	-0.412 to 0.019	NA	-0.015 to 0.0003	NA	NA
Experiment 3	-0.423 to -0.038	-0.418 to -0.024	NA	-0.020 to -0.002	NA	NA
Experiment 3b	-0.031 to 0.081	-0.337 to 0.072	NA	-0.014 to 0.004	NA	NA
H2: affPE ~ policy						
Experiment 1	-0.015 to 0.046	-0.015 to 0.047	NA	NA	NA	NA
Experiment 1b	-0.002 to 0.062	-0.001 to 0.062	NA	NA	NA	NA

H3a: persistence ~ affPE						
Experiment 1	-0.026 to 0.081	-0.030 to 0.077ǂ	-0.037 to 0.068	-0.004 to 0.007	-0.148 to -0.026 (OA: + assoc.)*	0.095 to 0.211 (PA: + assoc.)*
Experiment 1b	-0.136 to 0.011	-0.029 to -0.029ǂ	-0.175 to -0.045	-0.018 to -0.005	-0.279 to -0.123 (OA: + assoc.)*	0.053 to 0.218 (PA: + assoc)*
Experiment 2	0.072 to 0.186	0.075 to 0.184ǂ	0.059 to 0.163	0.005 to 0.015	-0.057 to 0.068 (OA: + assoc.)*	0.201 to 0.331 (PA: + assoc.)*
Experiment 3	0.034 to 0.136	0.040 to 0.142ǂ	0.021 to 0.123	0.002 to 0.010	-0.055 to 0.061 (OA: + assoc.)*	0.117 to 0.233 (PA: + assoc.)*
H3b: persistence ~ affPE x policy type						
Experiment 1	-0.126 to 0.104	-0.137 to 0.078ǂ	-0.126 to 0.068	-0.013 to 0.007	-0.390 to -0.142 (OA: no assoc.)	0.032 to 0.052 (PA: + assoc.)*
Experiment 1b	-0.092 to 0.213	-0.090 to 0.214	-0.094 to 0.165	-0.010 to 0.166	-0.200 to 0.133 (OA: + assoc.)*	0.020 to 0.045 (PA: + assoc.)*
Experiment 2	0.019 to 0.268	0.043 to 0.063ǂ	-0.013 to 0.183	-0.001 to 0.017	-0.294 to -0.026 (OA: + assoc.)*	0.090 to 0.342 (PA: + assoc.)*
Experiment 3	0.051 to 0.274	-0.004 to 0.205ǂ	0.070 to 0.260	0.006 to 0.022	-0.228 to 0.013 (OA: + assoc.)*	0.044 to 0.062 (PA: + assoc.)
H4a: escalation ~ affPE						
Experiment 1	0.010 to 0.032	0.010 to 0.032	0.023 to 0.045	0.002 to 0.005	0.016 to 0.042 (OA: - assoc.)	0.00008 to 0.025 (PA: - assoc.)
Experiment 1b	0.011 to 0.049	0.011 to 0.049	0.020 to 0.057	0.002 to 0.006	0.069 to 0.114 (OA: - assoc.)	-0.041 to 0.001 (PA: - assoc.)*
Experiment 2 (stealing only)	0.066 to 0.132	0.066 to 0.132	0.068 to 0.143	0.006 to 0.013	0.004 to 0.081 (OA: + assoc.)*	0.121 to 0.198 (PA: + assoc.)
Experiment 3 (stealing only)	0.060 to 0.131	0.061 to 0.132	0.086 to 0.163	0.007 to 0.014	-0.009 to 0.072 (OA: + assoc.)*	0.107 to 0.189 (PA: + assoc.)
Experiment 3b (destruction only)	-0.011 to 0.035	-0.011 to 0.035	0.009 to 0.053	0.0008 to 0.005	-0.072 to -0.018 (OA: + assoc.)*	0.028 to 0.080 (PA: + assoc.)*
H4b: escalation ~ affPE x policy type						
Experiment 1	-0.040 to 0.011	-0.002 to 0.004	-0.015 to 0.035	-0.002 to 0.004	-0.041 to 0.016 (OA: no assoc.)	-0.054 to 0.004 (PA: no assoc.)
Experiment 1b	-0.081 to 0.008	-0.006 to 0.002	-0.062 to 0.021	-0.006 to 0.002	-0.151 to -0.047 (OA: + assoc.)	-0.054 to 0.046 (PA: + assoc.)*

H5: escalation ~ affPE x induction x target ID						
Experiment 3	-0.143 to -0.003	-0.145 to -0.005	-0.165 to -0.016	-0.014 to -0.001	-0.167 to -0.005 (OA: no assoc.)*	-0.121 to 0.040 (PA: no assoc.)
(simple slope)	-0.074 to 0.025	-0.076 to 0.025	-0.119 to -0.010	-0.010 to -0.001	-0.085 to 0.028 (OA: no assoc.)*	-0.070 to 0.046 (PA: no assoc.)
Experiment 3b	-0.079 to 0.016	-0.079 to 0.016	-0.082 to 0.004	-0.008 to 0.0004	-0.119 to -0.010 (OA: + assoc.)*	-0.025 to 0.045 (PA: no assoc.)
(simple slope)	-0.082 to -0.010	-0.082 to -0.010	-0.070 to -0.007	-0.007 to -0.001	-0.129 to -0.047 (OA: - assoc.)	-0.070 to 0.012 (PA: no assoc.)

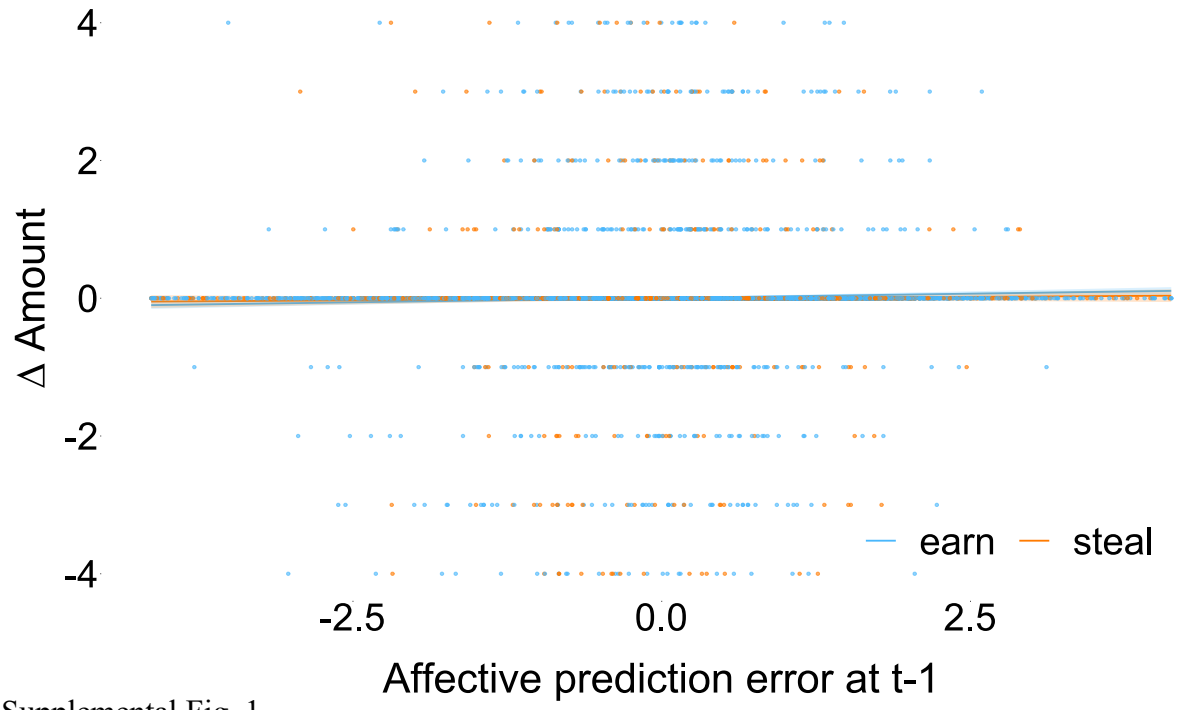
Supplemental Table 2.

Overview of 95% confidence intervals of regression weights corresponding to key pre-registered effects reported in the manuscript (all rows, column 2), supplemental robustness checks of those effects (columns 3-5), and models adjusting for predicted and post-outcome affect (columns 6 and 7). The "ɬ" symbol denotes convergence issues during model fitting. Statistical significance, which can be derived from confidence intervals excluding 0, is not highlighted, because the aim of this table is to document consistency of the results across model specifications. Highlighted cells and confidence intervals indicate directional deviations from findings reported in the manuscript, with yellow indicating findings that were not statistically significant in the manuscript and orange indicating findings that were significant in the manuscript. Notably, as for the robustness checks, confidence intervals are overwhelmingly consistent within each row, thereby indicating robustness to model specifications. One exception is evidence related to H3a in Experiment 1b, which counterintuitively—and perhaps spuriously—suggests that participants were less likely to persist in actions that recently felt better than expected. Another notable exception is evidence related to H5 in Experiment 3: The analytical approach specified in our pre-registration reveals a significant interaction in Experiment 3 and a significant simple slope in Experiment 3b, but other reasonable approaches (columns 4 and 5) would have additionally revealed a significant simple slope in Experiment 3.

The remaining two columns correspond to adjustment for predicted affect and post-outcome affect. The sign of the association between affective prediction errors and a given dependent variable often flipped depending on whether post-outcome affect or predicted affect were included in the prediction error model. This can be attributed to the inherent interdependence between the prediction error and its constituent parts, highlighting a key limitation of this analytical strategy. We included this strategy regardless in order to reflect the precedent in the literature. To address the problem of interdependence, we also include another strategy, which compares the fit of models including a given affective prediction error constituent (predicted affect or post-outcome affect) to a model including the composite affective prediction error. Cases where a constituent provided a better fit are denoted with a star. Evidently, when predicting persistence, individual constituent parts are overwhelmingly better suited to explain variance compared to prediction errors. When it comes to escalation, however, the picture is more mixed, showing evidence for the unique predictiveness of affective prediction errors regarding the key hypotheses reported in the manuscript (H4a and H5).

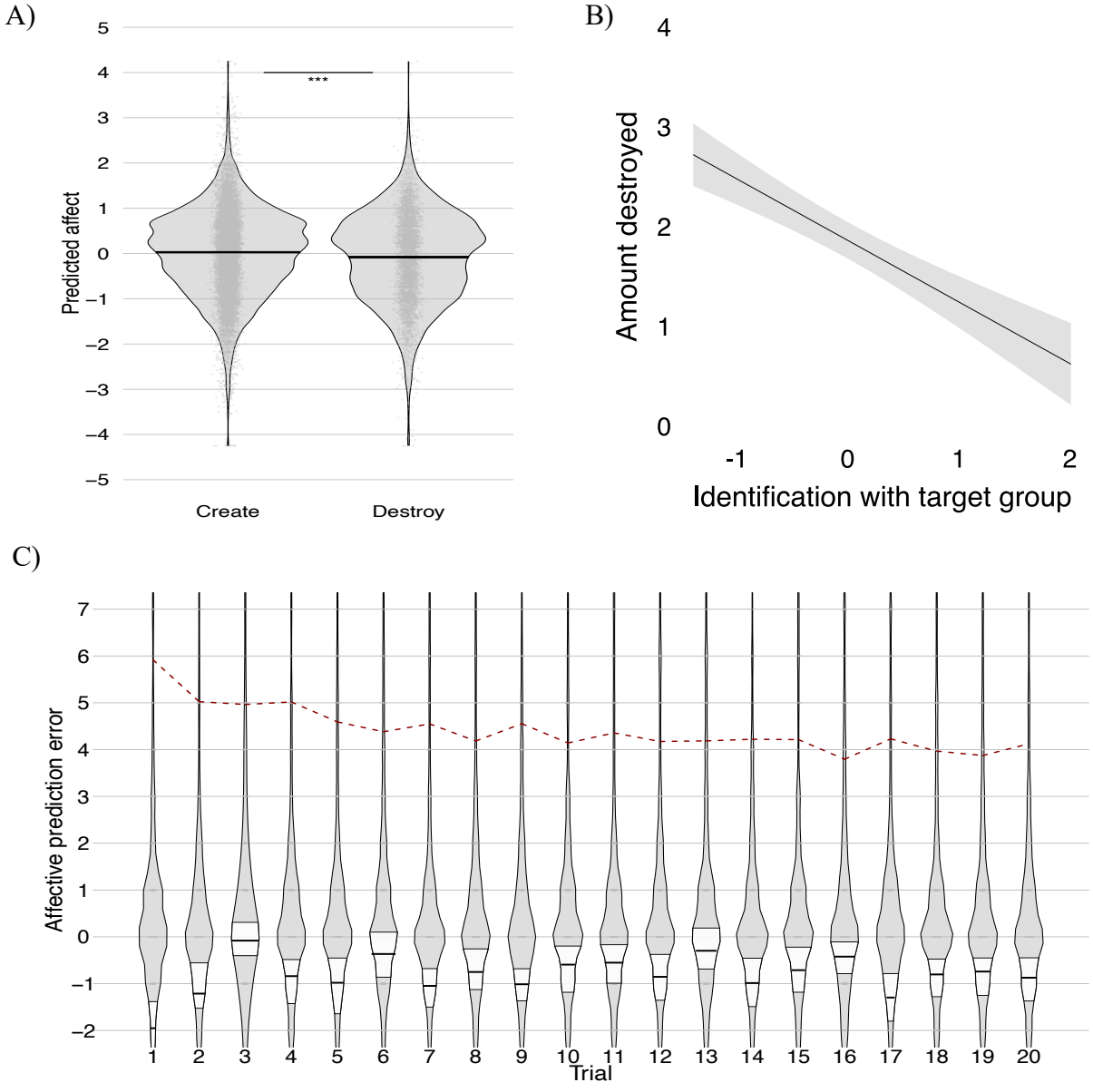
Supplemental Figures

Experiment 1 — Raw distribution of difference in amount chosen as a function of affective PEs by policy



Supplemental Fig. 1.
N = 901. Aggression escalation in Experiment 1: Model-based differences in amount chosen for same policy from t-1 to t (y axis) as a function of within-participant standardized affective prediction errors at t-1 (x axis). Separate lines are fitted for each prior policy. Shaded areas reflect 95% confidence intervals. Notably, in line with model assumptions, scaled residuals were normally distributed.

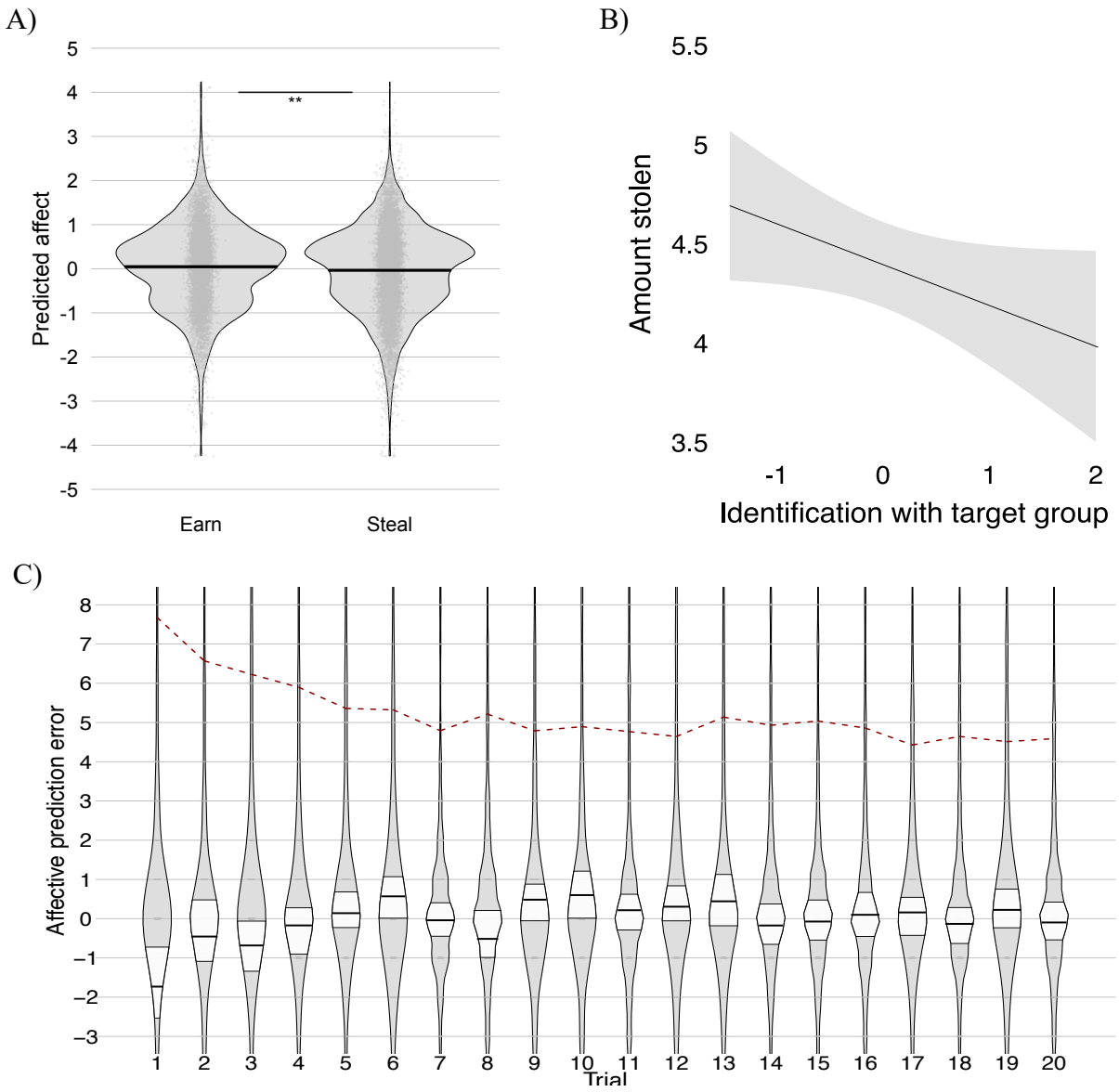
Experiment 1b — Predicted affect by policy; destruction by target ID, and affective PEs over time



Supplemental Fig. 2

N = 955. (A) Distributions of within-participant standardized affective predictions separated by policy (creating and destroying). Taking into account each participant's standards, participants expected destroying to feel worse than creating (β = -0.104; 95% CI, -0.132 to -0.056; η^2 = -0.09; n = 927; p < .001). (B) Model estimate of the relationship between standardized identification with the target's group on the x axis and the absolute amount destroyed from them on the y axis; participants destroyed fewer points from members of groups that they liked, valued, and felt connected to more. Shaded area reflects 95% CI. (C) Distributions of signed affective prediction errors across trials with mean absolute (unsigned) values shown in red. Signed values averaged below zero, reflecting systematic underestimations (b = -0.819; 95% CI, -1.061 to -0.577; η^2 = -0.08; n = 955; p < .0.001). As before, absolute values revealed decreasing prediction errors over time (b = -0.069; 95% CI, -0.088 to -0.051; η^2 = -0.01; n = 955; p < .001).

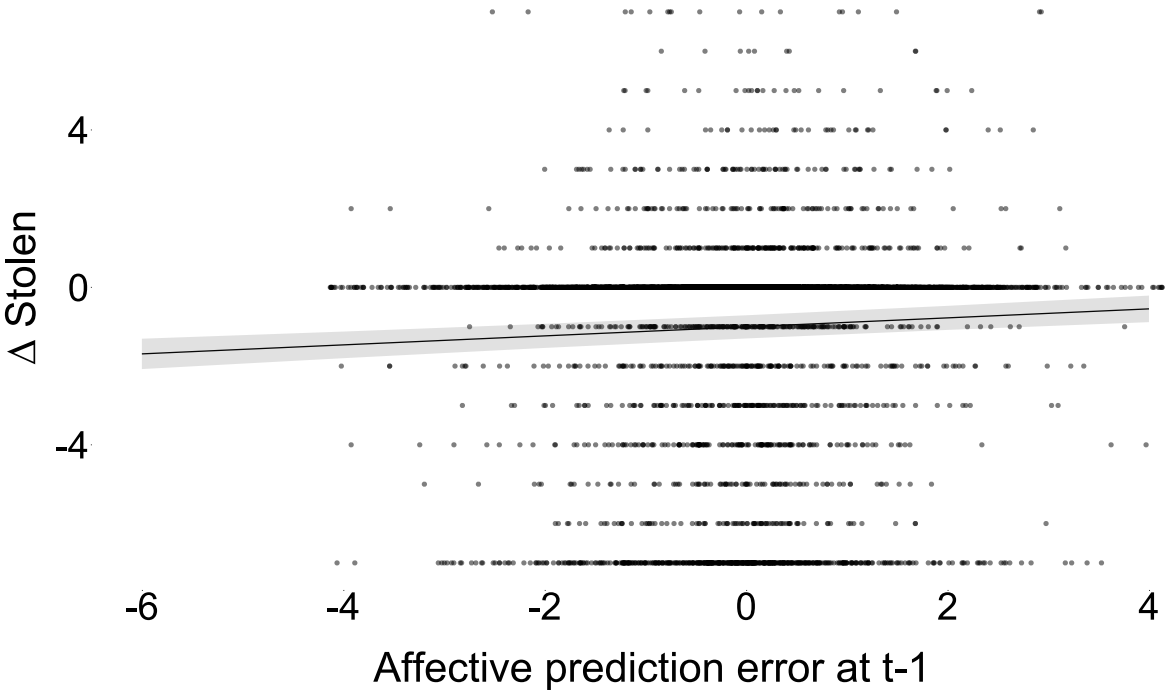
Experiment 2 — Predicted affect by policy; stealing by target ID, and affective PEs over time



Supplemental Fig. 3

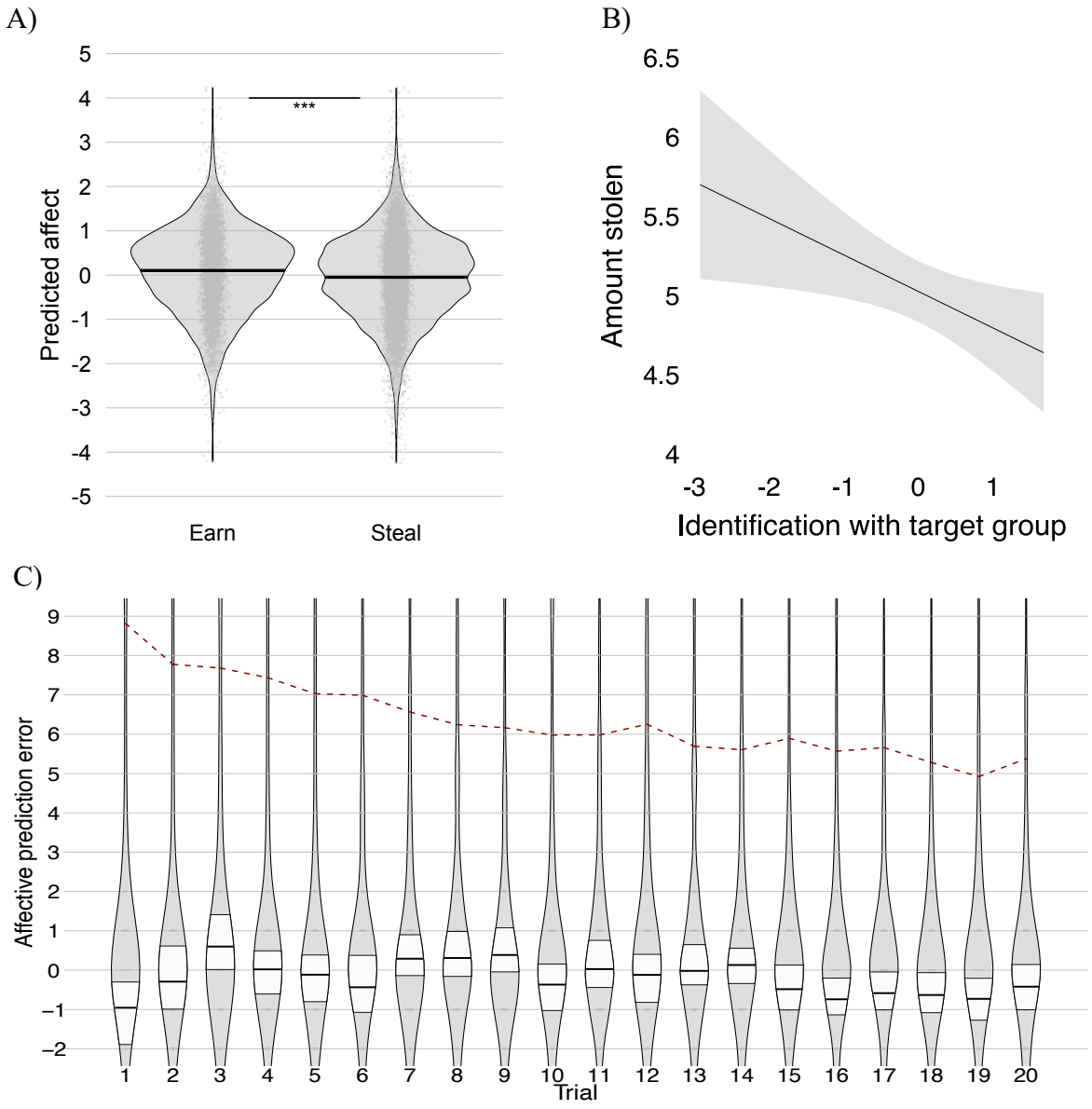
N = 942. (A) Distributions of within-participant standardized affective predictions separated by policy (earning and stealing). Taking into account each participant's standards, participants expected stealing to feel worse than earning (β = -0.087; 95% CI, -0.113 to -0.062; η^2 = -0.08; n = 932; $p < .001$). (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 fewer points from members of groups that they liked, valued, and felt connected to more. Shaded area reflects 95% CI. (C) Distributions of signed affective prediction errors across trials with mean absolute (unsigned) values shown in red. While signed values gravitate around zero, showing no overall imbalance between over- and underestimations (b = -0.042; 95% CI, -0.397 to 0.314; η^2 > -0.01; n = 942; p = .0.819), absolute values revealed decreasing prediction errors over time (b = -0.109; 95% CI, -0.128 to -0.089; η^2 = -0.01; n = 942; $p < .001$).

Experiment 2 —Raw distribution of difference in amount stolen as a function of affective PEs



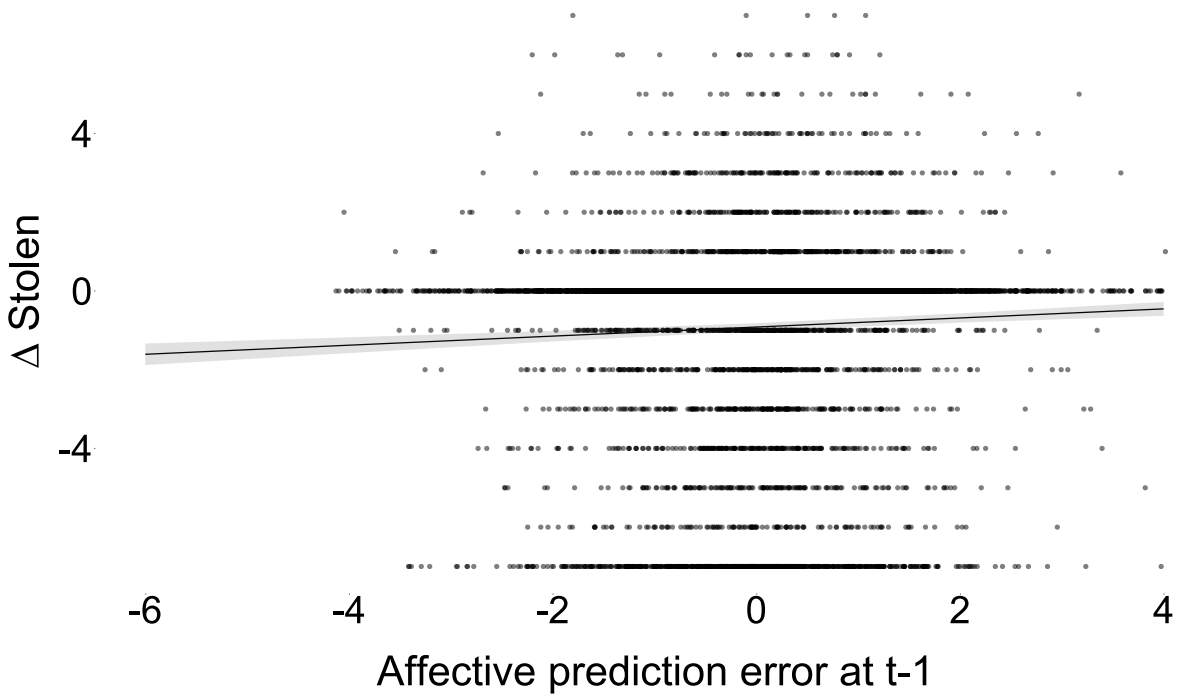
Supplemental Fig. 4.
N = 719. Aggression escalation in Experiment 2: Model-based differences in amount stolen from t-1 to t (y axis) as a function of within-participant standardized affective prediction errors at t-1 (x axis). Shaded areas reflect 95% confidence intervals.as a function of affective prediction errors at t-1. In line with model assumptions, scaled residuals were normally distributed.

Experiment 3 — Predicted affect by policy; stealing by target ID, and affective PEs over time



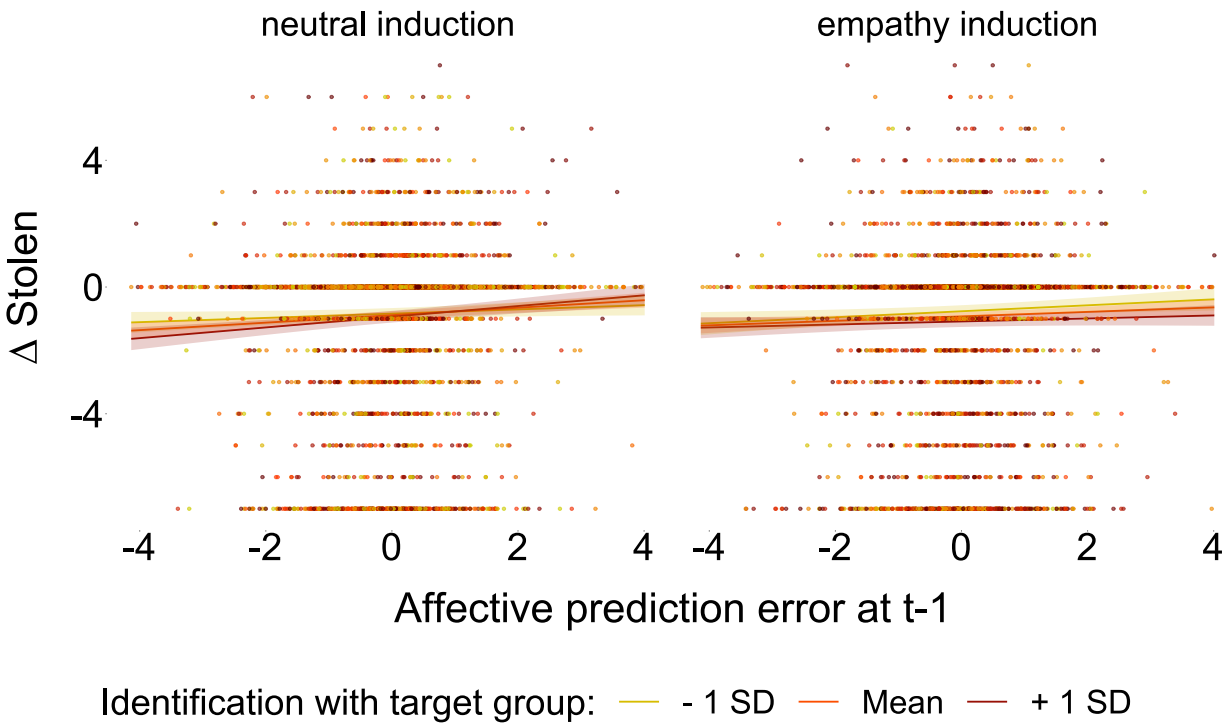
Supplemental Fig. 5.
N = 913. (A) Distributions of within-participant standardized affective predictions separated by policy (earning and stealing). Taking into account each participant's standards, participants expected stealing to feel worse than earning (β = -0.165; 95% CI, -0.192 to -0.138; η^2 = -0.15; n = 907; $p < .001$). (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 fewer points from members of groups that they liked, valued, and felt connected to more. Shaded area reflects 95% CI. (C) Distributions of signed affective prediction errors across trials with mean absolute (unsigned) values shown in red. While signed values gravitate around zero, showing no overall imbalance between over- and underestimations (b = -0.207; 95% CI, -0.594 to 0.180; η^2 = -0.02; n = 913; p = .294), absolute values revealed decreasing prediction errors over time (b = -0.157; 95% CI, -0.177 to -0.136; η^2 = -0.02; n = 913; $p < .001$).

Experiment 3 — Difference in amount stolen as a function of affective PEs



Supplemental Fig. 6.
N = 792. Aggression escalation in Experiment 3: Model-based differences in amount stolen from t-1 to t (y axis) as a function of within-participant standardized affective prediction errors at t-1 (x axis). Shaded areas reflect 95% confidence intervals. In line with model assumptions, scaled residuals were normally distributed.

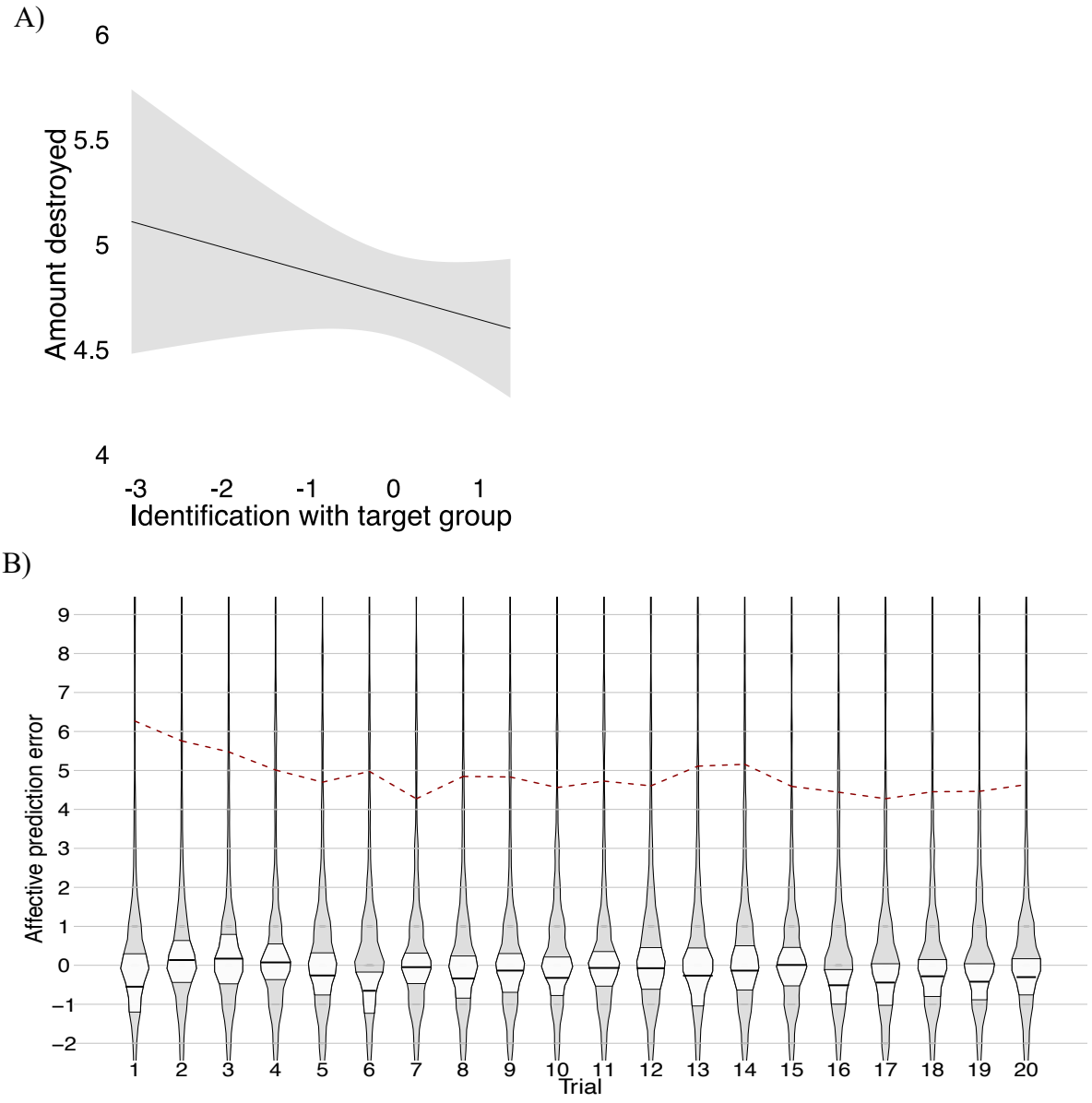
Experiment 3 —Difference in amount stolen as a function of affective PEs by target identification and emotion induction



Supplemental Fig.7.

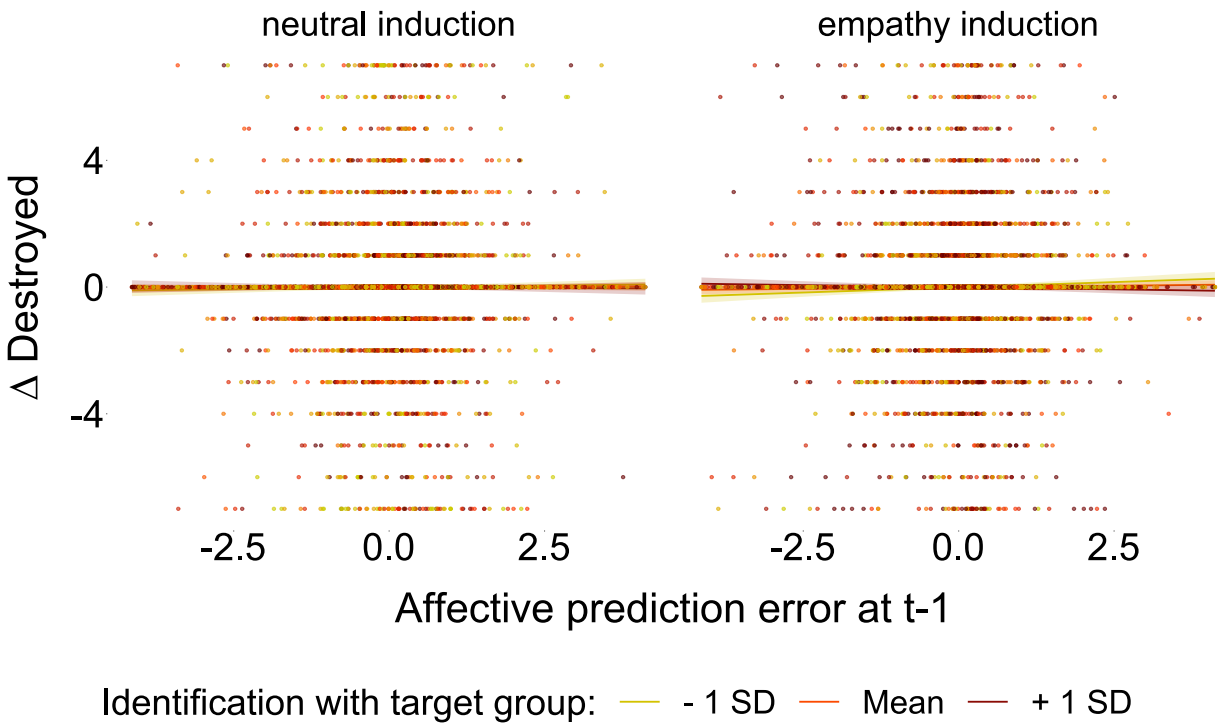
N = 792. Effects of empathy induction in Experiment 3: Model-based differences in amount stolen from t-1 to t (y axis) as a function of within-participant standardized affective prediction errors at t-1 (x axis), target identification (-1SD, yellow; mean, orange; +1SD, purple) and emotion induction condition (control condition, left panel; empathy induction, right panel). Following the empathy induction, the association between stealing 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% confidence intervals. In line with model assumptions, scaled residuals were normally distributed.

Experiment 3b — Destruction by target ID and affective PEs over time



Supplemental Fig. 8.
N = 887. (A) Model estimate of the relationship between standardized identification with the target's group on the x axis and the absolute amount destroyed from them on the y axis; participants were destroyed fewer points from members of groups that they liked, valued, and felt connected to more. Shaded area reflects 95% CI. (B) Distributions of signed affective prediction errors across trials with mean absolute (unsigned) values shown in red. While signed values gravitated around zero, showing no overall imbalance between over- and underestimations (b = -0.221; 95% CI, -0.551 to 0.109; η^2 = -0.02; n = 887; p = .189), absolute values revealed decreasing prediction errors over time (b = -0.058; 95% CI, -0.077 to -0.039; η^2 = -0.01; n = 887; p < .001).

Experiment 3b —Difference in amount destroyed as a function of affective PEs by target identification and emotion induction



Supplemental Fig.9.
N = 868. Effects of empathy induction in Experiment 3b: Model-based differences in amount destroyed from t-1 to t (y axis) as a function of within-participant standardized affective prediction errors at t-1 (x axis), target identification (-1SD, yellow; mean, orange; +1SD, purple) 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% confidence intervals.