Research Article

Psychological Science
1–14


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976241228504
www.psychologicalscience.org/PS

People Endorse Harsher Policies in Principle Than in Practice: Asymmetric Beliefs About Which Errors to Prevent Versus Fix

Eitan D. Rude and Franklin Shaddy
University of California, Los Angeles – Anderson School of Management

Abstract

Countless policies are crafted with the intention of punishing all who do wrong or rewarding only those who do right. However, this requires accommodating certain mistakes: some who do not deserve to be punished might be, and some who deserve to be rewarded might not be. Six preregistered experiments ($N = 3{,}484$ U.S. adults) reveal that people are more willing to accept this trade-off in principle, before errors occur, than in practice, after errors occur. The result is an asymmetry such that for punishments, people believe it is more important to prevent false negatives (e.g., criminals escaping justice) than to fix them, and more important to fix false positives (e.g., wrongful convictions) than to prevent them. For rewards, people believe it is more important to prevent false positives (e.g., welfare fraud) than to fix them and more important to fix false negatives (e.g., improperly denied benefits) than to prevent them.

Keywords

fairness, moral judgment, judgment and decision making, past versus future, public policy, punishments and rewards, omission versus commission, open data, open materials, preregistered

Received 8/4/23; Revision accepted 1/3/24

Governments, firms, and countless other institutions frequently use various types of punishment and reward policies. However, no policy is perfect. Sometimes, those who deserve to be punished or rewarded are not (false negatives); at other times, those who do not deserve to be punished or rewarded are (false positives). Which errors do people believe are worse, when, and why?

In this research, we develop a generalizable framework describing preferences regarding these errors (see Table 1). Specifically, we find that preferences to address false positives versus false negatives vary along two dimensions: (a) whether errors are considered before or after they occur and (b) whether they pertain to punishments or rewards.

For example, suppose an insurance company decides to increase the premiums charged to unsafe drivers—a punishment. Two types of mistakes are possible: some safe drivers might have their premiums raised (false positives), whereas some unsafe drivers might not (false negatives). We find that people believe it is more important to prevent false negatives than to fix them and more important to fix false positives than to prevent them.

For rewards, the opposite pattern holds. For example, suppose instead that the same insurance company decides to reduce the premiums charged to safe drivers—a reward. Some unsafe drivers might have their premiums reduced (false positives), whereas some safe drivers might not (false negatives). Here, we find that people believe it is more important to prevent false positives than to fix them and more important to fix false negatives than to prevent them.

To help explain these patterns, we first note that for punishments and rewards alike, false positives and false negatives can either harm "good actors" (those who do not deserve to be punished but are, and those who

Corresponding Author:
Eitan Rude, University of California, Los Angeles – Anderson School of Management
Email: eitan.rude.phd@anderson.ucla.edu

deserve to be rewarded, but are not) or help "bad actors" (those who deserve to be punished but are not, and those who do not deserve to be rewarded, but are). This common denominator matters because we expect people to naturally relate more to good actors harmed than to bad actors helped (Chambers & Davis, 2012; Sedikides et al., 2003). In other words, people can more easily imagine themselves as someone who does not deserve to be punished or deserves to be rewarded (as opposed to someone who deserves to be punished or does not deserve to be rewarded). We therefore propose that the good actors harmed will be relatively more vivid than the bad actors helped.

Second, judgments about errors to prevent versus fix can be conceptualized as judgments about the future versus the past. Critically, past outcomes are more accessible and concrete than future outcomes (Caruso et al., 2008; Kane et al., 2012; Tversky & Kahneman, 1973; Van Boven & Ashworth, 2007). Therefore, we additionally propose that when errors have already occurred (i.e., when considering errors to fix), there will be larger differences in vividness between good actors and bad actors relative to when errors have not yet occurred (i.e., when considering errors to prevent). Even if the latter outcomes are certain, they are less accessible and concrete because they are unrealized (Small & Loewenstein, 2005).

To illustrate, consider again that for punishments (e.g., increasing premiums charged to unsafe drivers), two types of errors are possible: those who deserve to be punished might not be (false negatives) and those who do not deserve to be punished might be (false positives). After these errors occur, people can more easily imagine themselves as a safe driver who had their rates raised by mistake (i.e., as a good actor) than as an unsafe driver who did not (i.e., as a bad actor).

Statement of Relevance

Administering punishments and rewards inevitably requires resolving trade-offs between different types of errors, and there is often considerable debate about which are the most problematic. For example, some promote aggressive law-enforcement tactics (e.g., "tough-on-crime" policies) out of concern for false negatives, whereas others seek exoneration of the wrongfully convicted out of concern for false positives (e.g., support for The Innocence Project). This research develops a generalizable framework for understanding these beliefs. Specifically, we find that people are concerned with different errors when evaluating proposed policies (i.e., considering errors to prevent) than when evaluating existing policies (i.e., considering errors to fix), depending on whether such policies pertain to punishments or rewards. Consequently, framing a policy one way or another (e.g., describing affirmative action as rewarding the underrepresented or punishing the overrepresented) can similarly shift preferences. This research accordingly provides a novel theoretical lens for understanding real-world phenomena spanning political, managerial, and marketing contexts.

Table 1. An Asymmetry Between the Types of Errors People Prefer to Prevent Versus Fix

Punishments				
Error type	Prevent	Preference	Fix	Example
False positives	Those who do not deserve to be punished will be punished	<	Those who did not deserve to be punished were punished	More important to fix wrongful convictions than to prevent these mistakes from happening in the first place
False negatives	Those who deserve to be punished will not be punished	>	Those who deserved to be punished were not punished	More important to prevent criminals from escaping justice than to fix these mistakes after the fact
Rewards				
Error type	Prevent	Preference	Fix	Example
False positives	Those who do not deserve to be rewarded will be rewarded	>	Those who did not deserve to be rewarded were rewarded	More important to prevent welfare fraud than to fix these mistakes after the fact
False negatives	Those who deserve to be rewarded will not be rewarded	<	Those who deserved to be rewarded were not rewarded	More important to fix improperly denied benefits than to prevent these mistakes from happening in the first place

Similarly, for rewards (e.g., reducing premiums for safe drivers), two types of errors are possible: those who deserve to be rewarded might not be (false negatives) and those who do not deserve to be rewarded might be (false positives). After these errors occur, people can more easily imagine themselves as a safe driver who did not have their rates reduced by mistake (i.e., as a good actor) than as an unsafe driver who did (i.e., as a bad actor). However, in both cases, before these errors occur, these same unrealized outcomes are less vivid.

Altogether, the hypothesized differences in vividness led us to predict that the most concerning types of errors will be those that both (a) harm good actors and (b) have already happened. Indeed, we find that people maintain stronger preferences for fixing false-positive punishments than for preventing them and stronger preferences for fixing false-negative rewards than for preventing them. The overall effect is the endorsement of harsher policies in principle (before errors occur, when people focus more on preventing mistakes that help bad actors) than in practice (after errors occur, when people focus relatively more on fixing mistakes that harm good actors).

Open Practices Statement

We report 14 preregistered experiments (six in the main text, eight in the appendix available online; total N = 7,278; see Table 2). All sample sizes were set a priori and were sufficient to detect a small interaction (f^2 = 0.06) with 80% power. This threshold was set on the basis of a pilot experiment that revealed a small interaction (f^2 = 0.06, 95% confidence interval, or CI = [0.02, 0.12]) and a subsequent power analysis that suggested that at least 47 participants per cell would be required to detect it. We therefore conservatively targeted a sample size of at least 50 participants per cell across all experiments.

Further, we disclose all measures, manipulations, and exclusions. All preregistrations, materials, data, and code are available at https://osf.io/69rjm/.

Finally, all experiments met the ethical requirements and legal guidelines of the University of California, Los Angeles's Institutional Review Board.

Experiment 1

Method

Participants. We recruited 300 participants from the behavioral laboratory at the University of California, Los Angeles's Anderson School of Management. We excluded participants who did not complete the survey in its entirety, yielding a final sample of 296 (age: M = 22.60 years, SD = 5.29 years; gender: 73% female, 25% male, 2% nonbinary).

Procedure. Experiment 1 employed a 2 (policy: punishment vs. reward) × 2 (frame: prevent vs. fix) between-subjects design.

For the punishment policy, participants read one of the following prompts:

1. **Punishments-Prevent:** "A policy is being designed to punish people. It will result in two mistakes. Only one of these mistakes can be prevented."
2. **Punishments-Fix:** "A policy was designed to punish people. It resulted in two mistakes. Only one of these mistakes can be fixed."

We then asked: "Which mistake should be [prevented/fixed]?" Participants selected between "10 individuals [will be/were] punished, but they [will not deserve it/did not deserve to be]" (false positives) and "10 individuals [will deserve/deserved] to be punished, but they [will not be/were not]" (false negatives).

For the reward policy, participants read one of the following prompts:

3. **Rewards-Prevent:** "A policy is being designed to reward people. It will result in two mistakes. Only one of these mistakes can be prevented."
4. **Rewards-Fix:** "A policy was designed to reward people. It resulted in two mistakes. Only one of these mistakes can be fixed."

We then asked: "Which mistake should be [prevented/fixed]?" Participants selected between "10 individuals [will be/were] rewarded, but they [will not deserve it/did not deserve to be]" (false positives) and "10 individuals [will deserve/deserved] to be rewarded, but they [will not be/were not]" (false negatives).

The prevent and fix frames were thus identical across the punishment and reward policies save for references to "punish" or "punished" and "reward" or "rewarded," respectively. There were no other differences across conditions. We counterbalanced the order of all choices.

Results

We coded false-negative choices as 1 and false-positive choices as 0, and we contrast-coded both policy (–1 for punishments; +1 for rewards) and frame (–1 for prevent; +1 for fix).

A logistic regression revealed a two-way policy-frame interaction, b = 0.65, 95% CI = [0.40, 0.91], z = 5.06, p < .001, odds ratio (OR) = 1.92, 95% CI = [1.50, 2.49] (see Fig. 1). Among those evaluating the punishment policy, a larger proportion elected to prevent false negatives (0.60, 95% CI = [0.48, 0.71]) than to fix them (0.29, 95% CI = [0.19, 0.40]; b = –0.63, 95% CI = [–0.97, –0.29], z = –3.66, p < .001, OR = 0.53, 95% CI = [0.38, 0.75]). Among those evaluating the reward

Table 2. Overview of Experiments

Experiment	*N*	Purpose	AsPredicted	Main finding	Proportion addressing false negatives (*N* per cell)				Significance
					Punishments		Rewards		
					Prevent	Fix	Prevent	Fix	
1	296	Basic effect (generic stimuli)	#130889	Participants were more likely to fix false-positive punishments than to prevent them and more likely to fix false-negative rewards than to prevent them.	0.60 (*N* = 72)	0.29 (*N* = 78)	0.54 (*N* = 74)	0.82 (*N* = 72)	***
2	917	Basic effect (naturalistic stimuli)	#79213	Conceptual replication across different scenarios: (a) a firm that docked pay/issued bonuses for poor/good performance, (b) an auto insurer that raised/reduced premiums for unsafe/safe driving, and (c) a municipality that assessed tax penalties/issued tax credits for wasting/conserving water.	0.53 (*N* = 234)	0.13 (*N* = 221)	0.43 (*N* = 218)	0.86 (*N* = 244)	***
3	983	Basic effect (framing)	#111192	Merely *framing* a policy (e.g., affirmative action) as a punishment or a reward shifted beliefs about which types of errors to prevent and fix.	0.25 (*N* = 249)	0.14 (*N* = 241)	0.60 (*N* = 240)	0.72 (*N* = 253)	***
4	565	Mediation	#98171	"Good actors" (i.e., those who did not deserve to be punished, but were, and those who deserve to be rewarded, but were not) were relatively more vivid than "bad actors" (i.e., those who deserve to be punished, but were not, and those who did not deserve to be rewarded, but were) when fixing versus preventing errors, and these perceptions statistically mediated the effect.	0.50 (*N* = 125)	0.09 (*N* = 160)	0.48 (*N* = 149)	0.89 (*N* = 131)	***
5a–b	723	Boundary condition	#93609 (a) #93459 (b)	The effect attenuated when a program was designed to measure water use, yielding false positives and false negatives, but no good actors and bad actors, as opposed to when a program was designed to *motivate* reduced water use (via punishments or rewards).	0.55 (*N* = 82)	0.14 (*N* = 102)	0.57 (*N* = 103)	0.83 (*N* = 76)	***

Note: *N*s per cell for Experiments 5a and 5b exclude the no motivation conditions.
***$p < .001$.

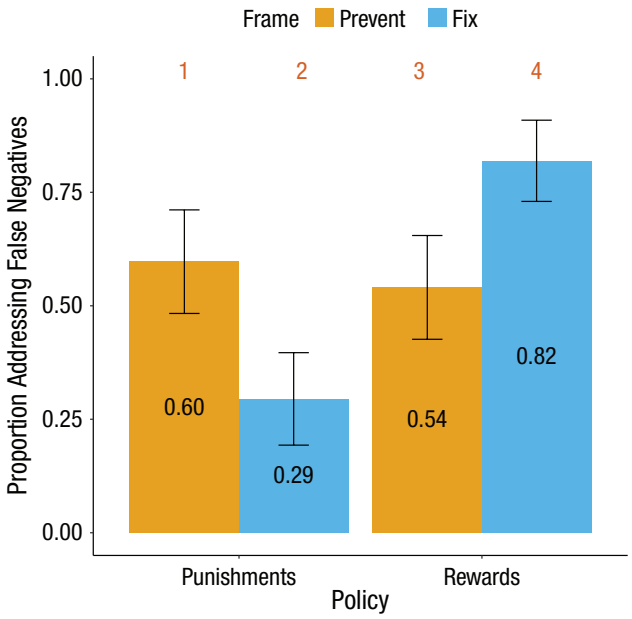

Fig. 1. Results from Experiment 1. For punishments, participants believed it was more important to prevent false negatives than to fix them, and more important to fix false positives than to prevent them; for rewards, participants believed it was more important to prevent false positives than to fix them, and more important to fix false negatives than to prevent them. Numbers above each bar correspond to the stimuli outlined in the procedure; error bars represent 95% confidence intervals.

policy, a larger proportion elected to fix false negatives (0.82, 95% CI = [0.73, 0.91]) than to prevent them (0.54, 95% CI = [0.43, 0.65]; b = 0.68, 95% CI = [0.30, 1.05], z = 3.51, $p < .001$, OR = 1.97, 95% CI = [1.35, 2.86]).

Discussion

Experiment 1 revealed the predicted asymmetry. Additionally, in a supplemental experiment that was otherwise identical to Experiment 1, we replicated this asymmetry when treating frame as a within-subjects factor (i.e., participants indicated which errors to both prevent and fix; see Supplemental Experiment 1 in the appendix available online). Experiment 2 accordingly tests more naturalistic scenarios.

Experiment 2

Method

Participants. We opened an Amazon Mechanical Turk Human Intelligence Task (MTurk HIT) for 1,000 participants. We excluded participants who failed a preregistered attention check, yielding a final sample of 917 (age: M = 42.25 years, SD = 12.94 years; gender: 56% female, 44% male, 1% nonbinary).

Procedure. Experiment 2 employed a 2 (policy: punishment vs. reward; between-subjects) × 2 (frame: prevent vs. fix; between-subjects) × 3 (scenario: pay vs. insurance vs. taxes; within-subjects) mixed design. We tested multiple scenarios to bolster generalizability but did not predict any systematic differences across them.

Each participant responded to three scenarios on three separate pages (see Table 3 and Table 4). For the punishment policy, the first scenario (pay) described docking pay for poor work performance, the second scenario (insurance) described assessing surcharges for unsafe driving, and the third scenario (taxes) described levying fines for using too much water during a drought. For the reward policy, the first scenario (pay) described issuing bonuses for good work performance, the second scenario (insurance) described the provision of discounts for safe driving, and the third scenario (taxes) described the issuance of tax credits for conserving water during a drought. As in Experiment 1, we described false positives and false negatives, and participants indicated which type of error should be prevented or fixed. We counterbalanced the order of all choices.

Results

We coded false-negative choices as 1 and false-positive choices as 0, and we contrast-coded both policy (–1 for punishments; +1 for rewards) and frame (–1 for prevent; +1 for fix). We then took the mean of each participant's choices across the three within-subjects scenarios. Note that this analysis deviated from our preregistration, which called for a mixed model. However, because all simple effects and two-way policy-frame interactions were significant and consistent with our predictions, we elected to present this simpler analysis (i.e., collapsing our results over scenario).

An ordinary least squares regression revealed a two-way policy-frame interaction, b = 0.21, 95% CI = [0.19, 0.23], $t(913)$ = 18.42, $p < .001$, f^2 = 0.37, 95% CI = [0.29, 0.46] (see Fig. 2). For those evaluating punishment policies, a larger proportion of participants elected to prevent false negatives (0.53, 95% CI = [0.48, 0.59]) than to fix them (0.13, 95% CI = [0.10, 0.16]; b = –0.20, 95% CI = [–0.23, –0.17], $t(453)$ = –12.50, $p < .001$, d = 1.17, 95% CI = [0.97, 1.37]). For those evaluating reward policies, a larger proportion of participants elected to fix false negatives (0.86, 95% CI = [0.82, 0.90]) than to prevent them (0.43, 95% CI = [0.37, 0.48]; b = 0.22, 95% CI = [0.19, 0.25], $t(460)$ = 13.55, $p < .001$, d = 1.26, 95% CI = [1.06, 1.46]).

Discussion

Experiment 2 conceptually replicated Experiment 1 with richer stimuli, bolstering generalizability. In

Table 3. Experiment 2 Stimuli

	Punishments		Rewards	
Scenario	Prevent	Fix	Prevent	Fix
Pay	XYZ Inc. is planning to dock the pay of 100 salespeople by $500 as a penalty for their low performance.	XYZ Inc. docked the pay of 100 salespeople by $500 as a penalty for their low performance.	XYZ Inc. is planning to award 100 salespeople a bonus payment of $500 in recognition of their high performance.	XYZ Inc. awarded 100 salespeople a bonus payment of $500 in recognition of their high performance.
Insurance	ABC Auto Insurance Co. is conducting an audit of their policyholders and plans to increase monthly premium payments by $50 for 100 of their policyholders who are deemed to be the most reckless drivers.	ABC Auto Insurance Co. conducted an audit of their policyholders and decided to increase monthly premium payments by $50 for 100 of their policyholders who were deemed to be the most reckless drivers.	ABC Auto Insurance Co. is conducting an audit of their policyholders and plans to reduce monthly premium payments by $50 for 100 of their policyholders who are deemed to be the safest drivers.	ABC Auto Insurance Co. conducted an audit of their policyholders and decided to reduce monthly premium payments by $50 for 100 of their policyholders who were deemed to be the safest drivers.
Taxes	Smallville is experiencing a drought, and as a result each household is being given an allotment for how much water they can consume over the course of a year. Annual bills will soon be computed, and the city's Superintendent of Water plans to assess a fine of $250 upon 100 residents who go too far over their allotment.	Smallville is experiencing a drought, and as a result each household was given an allotment for how much water they could consume over the course of a year. Annual bills were recently computed, and the city's Superintendent of Water assessed a fine of $250 upon 100 residents who went too far over their allotment.	Smallville is experiencing a drought, and as a result each household is being given an allotment for how much water they can consume over the course of a year. Annual bills will soon be computed, and the city's Superintendent of Water plans to award a credit of $250 to 100 residents who remain sufficiently below their allotment.	Smallville is experiencing a drought, and as a result each household was given an allotment for how much water they could consume over the course of a year. Annual bills were recently computed, and the city's Superintendent of Water awarded a credit of $250 to 100 residents who remained sufficiently below their allotment.

Experiment 3, we tested whether merely framing a policy as a punishment or reward shifts beliefs similarly.

Experiment 3

We described affirmative action as either rewarding the underrepresented or punishing the overrepresented. Past research has similarly framed these policies as either helping minority or harming majority individuals and groups (Crosby et al., 2003; Lowery et al., 2006; Munguia Gomez & Levine, 2022).

Method

Participants. We requested 1,000 participants on Prolific Academic. We excluded participants who failed a preregistered attention check, yielding a final sample of 983 (age: $M = 39.76$ years, $SD = 14.18$ years; gender: 48% female, 50% male, 1% nonbinary).

Procedure. Experiment 3 employed a 2 (policy: punishment vs. reward) × 2 (frame: prevent vs. fix) between-subjects design.

For the punishment policy, participants read, "An organization [is planning to implement/has implemented] an affirmative action policy to punish people from over-represented backgrounds. However, this policy [will result/has resulted] in two types of mistakes." We then asked, "Which mistake should be [prevented/fixed]?" Participants selected between false negatives ("Some people from over-represented backgrounds [will not be/were not] punished (even though they [will deserve/deserved] to be punished)") and false positives ("Some people from under-represented backgrounds [will be/were] punished (even though they [will not/did not] deserve to be punished)").

For the reward policy, participants read: "An organization [is planning to implement/has implemented] an affirmative action policy to reward people from underrepresented backgrounds. However, this policy [will result/has resulted] in two types of mistakes." We then asked: "Which mistake should be [prevented/fixed]?" Participants selected between false negatives ("Some people from under-represented backgrounds [will not be/were not] rewarded (even though they [will deserve/deserved] to be rewarded)") and false positives ("Some people from

Table 4. Experiment 2 Choices

	Punishments		Rewards	
Scenario	Prevent	Fix	Prevent	Fix
Pay	Ten of the 100 employees whose pay will be docked will not deserve it (that is, they will not in fact be low performers). [FP] Ten additional employees will deserve to have their pay docked (that is, they will in fact be low performers), but will not. [FN]	Ten of the 100 employees whose pay was docked did not deserve it (that is, they were not in fact low performers). [FP] Ten additional employees deserved to have their pay docked (that is, they were in fact low performers), but did not. [FN]	Ten of the 100 employees who will receive a bonus will not deserve it (that is, they will not in fact be high performers). [FP] Ten additional employees will deserve a bonus (that is, they will in fact be high performers), but will not receive it. [FN]	Ten of the 100 employees who received a bonus did not deserve it (that is, they were not in fact high performers). [FP] Ten additional employees deserved a bonus (that is, they were in fact high performers), but did not receive it. [FN]
Insurance	Ten of the 100 policyholders whose premiums will be increased will not deserve it (that is, they will not in fact be reckless drivers). [FP] Ten additional policyholders will deserve to have their premiums increased (that is, they will in fact be reckless drivers) but will not. [FN]	Ten of the 100 policyholders whose premiums were increased did not deserve it (that is, they were not in fact reckless drivers). [FP] Ten additional policyholders deserved an increase in their premiums (that is, they were in fact reckless drivers), but did not receive it. [FN]	Ten of the 100 policyholders whose premiums will be reduced will not deserve it (that is, they will not in fact be safe drivers) [FP] Ten additional policyholders will deserve a reduction in their premiums (that is, they will in fact be safe drivers), but will not receive it. [FN]	Ten of the 100 policyholders whose premiums were reduced did not deserve it (that is, they were not in fact safe drivers). [FP] Ten additional policyholders deserved a reduction in their premiums (that is, they were in fact safe drivers), but did not receive it. [FN]
Taxes	Ten of the 100 residents who will be fined will not deserve it (that is, they will not have in fact gone over their allotment). [FP] Ten additional residents will deserve to be fined (that is, they will have in fact have gone over their allotment), but will not be. [FN]	Ten of the 100 residents who were fined did not deserve it (that is, they did not in fact go over their allotment). [FP] Ten additional residents deserved to be fined (that is, they did in fact go over their allotment), but were not. [FN]	Ten of the 100 residents who will receive a credit will not deserve it (that is, they will not have in fact remained below their allotment). [FP] Ten additional residents will deserve to receive a credit (that is, they will have in fact remained below their allotment), but will not. [FN]	Ten of the 100 residents who received a credit did not deserve it (that is, they did not in fact remain below their allotment). [FP] Ten additional residents deserved to receive a credit (that is, they did in fact remain below their allotment), but did not. [FN]

Note: FP = false positives; FN = false negatives. Designations were not shown to participants.

over-represented backgrounds [will be/were] rewarded (even though they [will not/did not] deserve to be rewarded)"). We counterbalanced the order of all choices.

Results

We coded false-negative choices as 1 and false-positive choices as 0, and we contrast-coded both policy (–1 for punishment; +1 for reward) and frame (–1 for prevent; +1 for fix).

A logistic regression revealed a two-way policy-frame interaction, $b = 0.30$, 95% CI = [0.15, 0.45], $z = 3.95$, $p < .001$, $OR = 1.35$, 95% CI = [1.16, 1.57] (see Fig. 3). When the affirmative action policy was described as a punishment, a larger proportion of participants elected to prevent false negatives (0.25, 95% CI = [0.20, 0.30]) than to fix them (0.14, 95% CI = [0.10, 0.19]; $b = –0.35$, 95% CI = [–0.58, –0.12], $z = –2.98$, $p = .003$, $OR = 0.70$, 95% CI = [0.56, 0.89]). When the same policy was described as a reward, a larger proportion of participants elected to fix false negatives (0.72, 95% CI = [0.66, 0.77]) than to prevent them (0.60, 95% CI = [0.54, 0.67]; $b = 0.25$, 95% CI = [0.06, 0.44], $z = 2.60$, $p = .009$, $OR = 1.28$, 95% CI = [1.06, 1.55]; see Fig. 3).

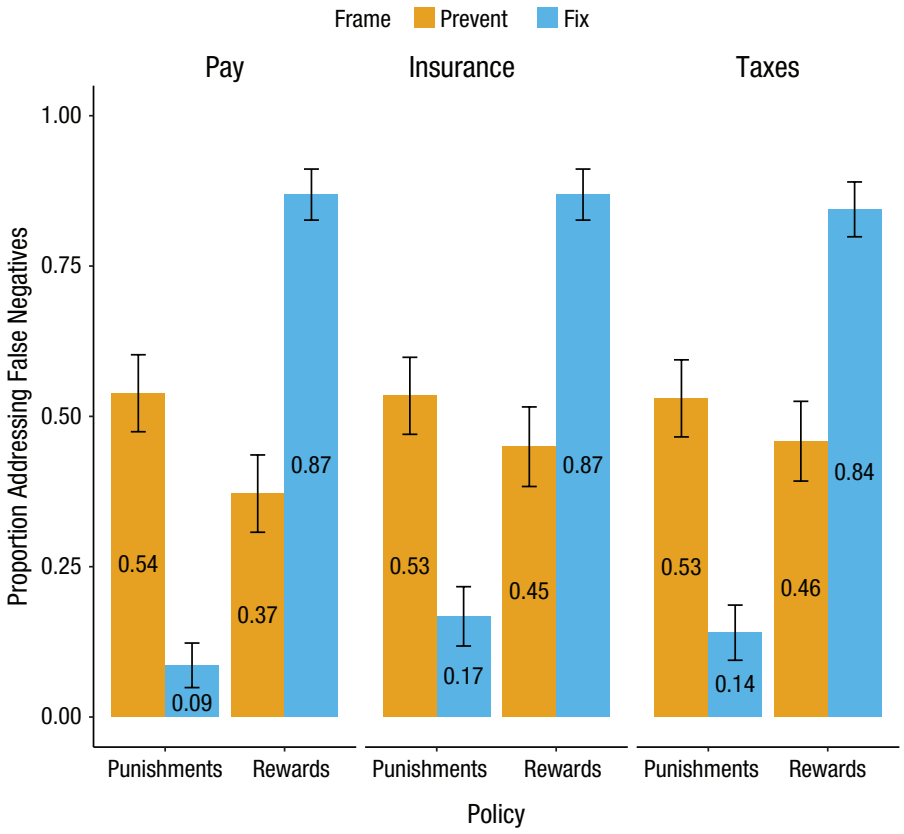

Fig. 2. Results from Experiment 2. Across the managerial context (pay), the consumer context (insurance), and the policy context (taxes), participants believed that for punishments it was more important to prevent false negatives than to fix them and more important to fix false positives than to prevent them; for rewards, participants believed it was more important to prevent false positives than to fix them and more important to fix false negatives than to prevent them. Error bars represent 95% confidence intervals.

Discussion

Experiments 1 through 3 offer convergent evidence for the basic effect. In a second supplemental experiment, we also elicited preferences via titration (as opposed to forced choice), and the resulting "exchange rates" conceptually replicated the results observed in Experiments 1 through 3 (see Supplemental Experiment 2 in the appendix available online). We therefore designed Experiment 4 to probe one potential mechanism: the relative vividness of good actors to bad actors.

Experiment 4

Method

Participants. We opened an MTurk HIT for 600 participants. We excluded participants who failed a preregistered attention check, yielding a final sample of 565 (age: M = 42.35 years, SD = 13.12 years; gender: 50% female, 49% male, 1% nonbinary).

Procedure. Experiment 4 employed a 2 (policy: punishment vs. reward) × 2 (frame: prevent vs. fix) between-subjects design.

Participants reviewed the pay scenario from Experiment 2. After indicating which type of error to prevent or fix, participants rated, on a separate page, the vividness of those affected by each type of error using measures adapted from Keller and Block (1997). Specifically, we asked, "How vivid are these salespeople?"; "How personal are the stories of these salespeople?"; "How concrete do these salespeople feel?"; "How easy is it to imagine these salespeople?"; "How easy is it to relate to these salespeople?"; and "How easy is it to picture these salespeople?" (each rated on a scale ranging from *not at all*, 1, to *extremely*, 7). We counterbalanced the order of all choices.

Results

We coded false-negative choices as 1 and false-positive choices as 0, and we contrast-coded both policy (–1 for

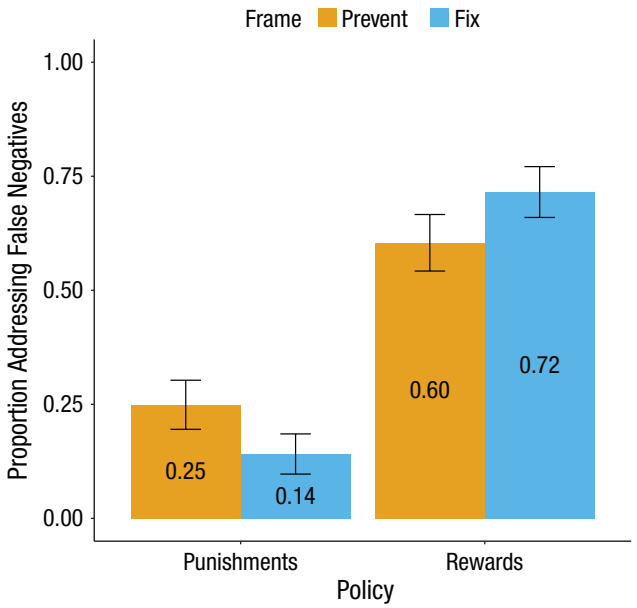

Fig. 3. Results from Experiment 3. Framing affirmative action as either "rewarding" the underrepresented or "punishing" the overrepresented yielded asymmetric beliefs about which types of errors should be prevented versus fixed. Error bars represent 95% confidence intervals.

punishment; +1 for reward) and frame (–1 for prevent; +1 for fix).

A logistic regression revealed a two-way policy-frame interaction (b = 1.11, 95% CI = [0.89, 1.34], z = 9.64, p < .001, OR = 3.04, 95% CI = [2.44, 3.83]), directly replicating Experiment 2. Among those evaluating the punishment policy, a larger proportion elected to prevent false negatives (0.50, 95% CI = [0.41, 0.58]) than to fix them (0.09, 95% CI = [0.04, 0.13]; b = –1.16, 95% CI = [–1.49, –0.84], z = –7.01, p < .001, OR = 0.31, 95% CI = [0.23, 0.43]). Among those evaluating the reward policy, a larger proportion elected to fix false negatives (0.89, 95% CI = [0.83, 0.94]) than to prevent them (0.48, 95% CI = [0.40, 0.56]; b = 1.06, 95% CI = [0.74, 1.37], z = 6.61, p < .001, OR = 2.89, 95% CI = [2.10, 3.94]).

Next, for each participant and for each type of error, we computed absolute vividness scores by averaging all six scale items (α = .92). Then, to construct a relative vividness score for each participant, we subtracted the absolute vividness score for the false positives from the absolute vividness score for the false negatives.

An ordinary least squares regression revealed a two-way policy-frame interaction, b = –0.14, 95% CI = [–0.23, –0.05], $t(561)$ = –3.12, p = .002, f^2 = 0.02, 95% CI = [0.002, 0.050]. For the punishment policy, false positives (i.e., the good actors, or the overperforming employees) were comparatively more vivid than false negatives (i.e., the bad actors, or the underperforming employees) in the fix frame (M = 0.76, 95% CI = [0.58, 0.93]) than in the prevent frame (M = 0.46, 95% CI = [0.27, 0.64]); b = 0.15, 95% CI = [0.02, 0.27], $t(283)$ = 2.31, p = .021, d = 0.27, 95% CI = [0.03, 0.50]). For the reward policy, false negatives (i.e., the good actors, or the overperforming employees) were comparatively more vivid than false positives (i.e., the bad actors, or the underperforming employees) in the fix frame (M = –0.95, 95% CI = [–1.12, –0.77]) than in the prevent frame (M = –0.68, 95% CI = [–0.85, –0.51]); b = –0.13, 95% CI = [–0.26, –0.01], $t(278)$ = –2.10, p = .036, d = 0.26, 95% CI = [0.02, 0.49]).

Finally, we tested whether relative vividness mediated the effect of frame on preferences to address false positives versus false negatives. We tested for mediation separately across the punishment and reward policies (with 10,000 bootstrapped samples for each), treating frame as the independent variable, choice as the dependent variable, and relative vividness as the mediator. For the punishment policy, choice was mediated by relative vividness (indirect effect: b = –0.008, p = .026, 95% CI = [–0.0198, –0.0007], proportion mediated = 0.05, 95% CI = [0.004, 0.127]). Similarly, for the reward policy, choice was mediated by relative vividness (indirect effect: b = 0.007, p = .042, 95% CI = [0.0002, 0.0187], proportion mediated = 0.04, 95% CI = [0.001, 0.114]). For the punishment and reward policies, sensitivity analyses (Imai et al., 2010) indicated that at ρs of –0.21 and –0.18, respectively, the average causal mediation effect was 0.00.

Discussion

Although there are numerous well-documented limitations associated with the use of statistical mediation to collect psychological process evidence (e.g., Imai et al., 2010; Rohrer et al., 2022), the results of Experiment 4 offer initial, suggestive evidence for the potential role of relative vividness. We comment on additional possible processes in the General Discussion. Our final experiments test a boundary condition.

Experiments 5a and 5b

If the relative vividness of good actors to bad actors helps explain, in part, different preferences regarding which errors to prevent or fix, then describing policies that generate false positives and false negatives without yielding corresponding good actors or bad actors should attenuate the effect. We thus manipulated whether a program was intended to motivate reduced water use (via punishments and rewards) or simply measure it.

Method

Participants. Given that all simple effects of frame (e.g., prevent vs. fix) within the punishment and reward conditions in Experiments 1 through 4 were significant (and in opposite directions), we tested punishments and rewards separately in Experiments 5a and 5b to maximize statistical power. For Experiment 5a, we opened an MTurk HIT for 400 participants. We excluded participants who failed a preregistered attention check, yielding a final sample of 360 (age: M = 40.77 years, SD = 12.06 years; gender: 55% female, 44% male, 1% nonbinary). For Experiment 5b, we opened an MTurk HIT for 400 workers. We excluded participants who failed a preregistered attention check, yielding a final sample of 363 (age: M = 40.69 years, SD = 14.04 years; gender: 50% female, 50% male, 0% nonbinary).

Procedure. Experiments 5a and 5b both employed a 2 (frame: prevent vs. fix) × 2 (goal: motivation vs. no motivation) between-subjects design and adapted the taxes scenario from Experiment 2.

Experiment 5a tested punishments, and Experiment 5b tested rewards. All participants in the motivation condition of each experiment first read, "Smallville [is experiencing/recently experienced] a drought, and as a result the local government [will be piloting/piloted] a program in which they [will test/tested] whether they [can/could] effectively encourage households to reduce their water use by collecting water usage information from 'smart meters.' Smallville's local government [plans to randomly select/randomly selected] 100 households to join the program based upon the last digit of their telephone number."

In the motivation condition of Experiment 5a, participants learned that "at the end of [the coming year/the year], the government [will assess/assessed] a fine of $500 upon any household which [increases/increased] their water use by 25% or more." Participants then chose to prevent or fix either false negatives ("10 households [will deserve/deserved] to be fined $500 but [will not be/were not]") or false positives ("10 of the households that [will be/were] fined $500 [will/did] not actually deserve it"). In the motivation condition of Experiment 5b, participants learned that "at the end of [the coming year/the year], the government [will issue/issued] a rebate of $500 to any household which [reduces/reduced] their water use by 25% or more." Participants then chose to prevent or fix either false negatives ("10 households [will deserve/deserved] to be issued a rebate of $500 but [will not be/were not]") or false positives ("10 of the households that [will be/were] issued a $500 rebate [will/did] not actually deserve it").

The no motivation conditions in Experiments 5a and 5b were identical. All participants first read: "Smallville [will be piloting/piloted] a program in which they [will test/tested] whether they [can/could] effectively collect water usage information from 'smart meters.' Smallville's local government [plans to randomly select/randomly selected] 100 households to join the program based upon the last digit of their telephone number." Participants then chose to prevent or fix either false negatives ("10 households should [be/have been] enrolled in the program but [will not be/were not]") or false positives ("10 of the households that [will be/were] enrolled in the program should not [be/have been]"). We counterbalanced the order of all choices in Experiments 5a and 5b.

Results

We coded false-negative choices as 1 and false-positive choices as 0, and we contrast-coded both frame (–1 for prevent; +1 for fix) and goal (–1 for no motivation; +1 for motivation).

For Experiment 5a, a logistic regression revealed a two-way frame-goal interaction, b = –0.60, 95% CI = [–0.84, –0.36], z = –4.91, p < .001, OR = 0.55, 95% CI = [0.43, 0.69] (see Fig. 4). In the motivation condition, there was a simple effect of frame (b = –1.00, 95% CI = [–1.37, –0.62], z = –5.22, p < .001, OR = 0.37, 95% CI = [0.25, 0.54]), conceptually replicating Experiment 2. Specifically, a larger proportion of participants elected to prevent false negatives (0.55, 95% CI = [0.46, 0.65]) than to fix them (0.14, 95% CI = [0.07, 0.22]). In the no motivation condition, there was no simple effect of frame (b = 0.20, 95% CI = [–0.10, 0.49], z = 1.32, p = .187, OR = 1.22, 95% CI = [0.90, 1.63]).

For Experiment 5b, a logistic regression revealed a two-way frame-goal interaction, b = 0.22, 95% CI = [–0.01, 0.45], z = 1.89, p = .059, OR = 1.24, 95% CI = [0.99, 1.56] (see Fig. 4). In the motivation condition, there was a simple effect of frame (b = 0.66, 95% CI = [0.32, 1.00], z = 3.79, p < .001, OR = 1.93, 95% CI = [1.38, 2.72]), conceptually replicating Experiment 2. Specifically, a larger proportion of participants elected to fix false negatives (0.83, 95% CI = [0.76, 0.91]) than to prevent them (0.57, 95% CI = [0.47, 0.68]). In the no motivation condition, there was no simple effect of frame (b = 0.22, 95% CI = [–0.08, 0.52], z = 1.43, p = .152, OR = 1.25, 95% CI = [0.92, 1.68]).

Discussion

Experiments 5a and 5b suggest that our account does not extend to any policy that generates false positives and false negatives, revealing an important boundary condition: absent good actors and bad actors who are erroneously harmed and helped, the effect is attenuated.

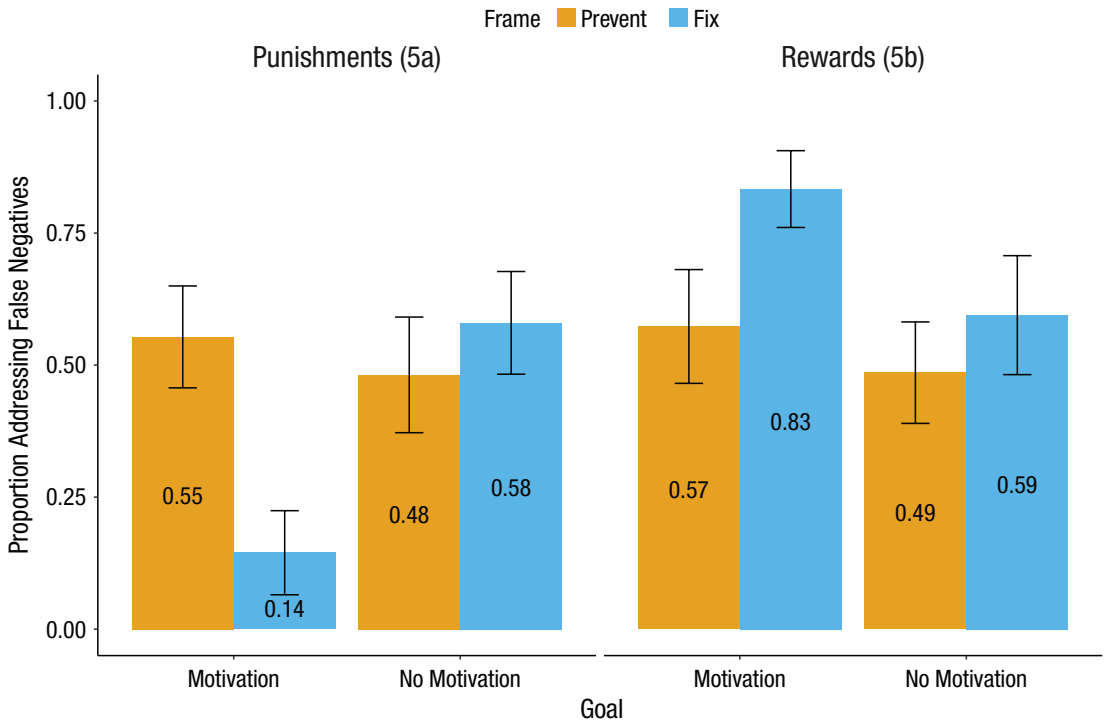

Fig. 4. Results from Experiments 5a and 5b. When a program that was intended to measure water use (as opposed to motivating reduced water use) yielded false positives and false negatives, but no good actors or bad actors, the effect attenuated. Error bars represent 95% confidence intervals.

General Discussion

When administering punishments and rewards, mistakes happen. Do some mistakes seem worse than others? This research finds that the answer depends not only on whether mistakes pertain to punishments or rewards but also on whether they are evaluated before or after they occur.

It thus offers a novel theoretical backdrop for understanding many real-world policy debates. For example, in a series of supplemental studies, participants chose between unspecified numbers of each type of error in the context of 10 real-world punishment and reward policies (see Supplemental Experiments 3 and 4 in the appendix available online). Even in this less controlled setting, we observed patterns directionally consistent with our laboratory experiments (though not all differences were significant; see Figs. 5 and 6). These findings potentially suggest that at least some disagreement about these issues may relate to how (or when) they are evaluated.

Notably, across experiments, the asymmetry seems to have been largely driven by beliefs about which errors to fix—that is, all choice shares for fixing, but not for preventing, differed significantly from 50%. Though we do not make normative claims regarding whether these preferences constitute a mistake, people may indeed have stronger convictions about which errors to fix than to prevent. If so, public support might be higher for policies that hew closer to the preferences observed in the fix frames. Additionally, it is unclear whether people themselves view these inconsistencies as problematic, given that we replicated these patterns when asking participants to indicate which errors to both prevent and fix simultaneously, using a within-subjects design (see Supplemental Experiment 1 in the appendix available online; Nielsen & Rehbeck, 2022).

Separately, false-positive and false-negative errors arise in numerous contexts, raising a natural question about generalizability. For example, the replication crisis fundamentally reflects a tension between the types of errors that the scientific enterprise has chosen to prevent versus fix (Ioannidis, 2005); medical testing requires calibrating tolerance for false-positive and false-negative results; markets that pick winners and losers sometimes err. Although the results of Experiments 5a and 5b suggest that these beliefs may be context dependent, it is still an open question whether they arise in other settings.

Several limitations warrant discussion and suggest other potential mechanisms. For example, inferences about the certainty of outcomes might have differed across our studies. We presented information unambiguously (e.g., "10 individuals will be punished, but they will not deserve it") to cleanly test our predictions,

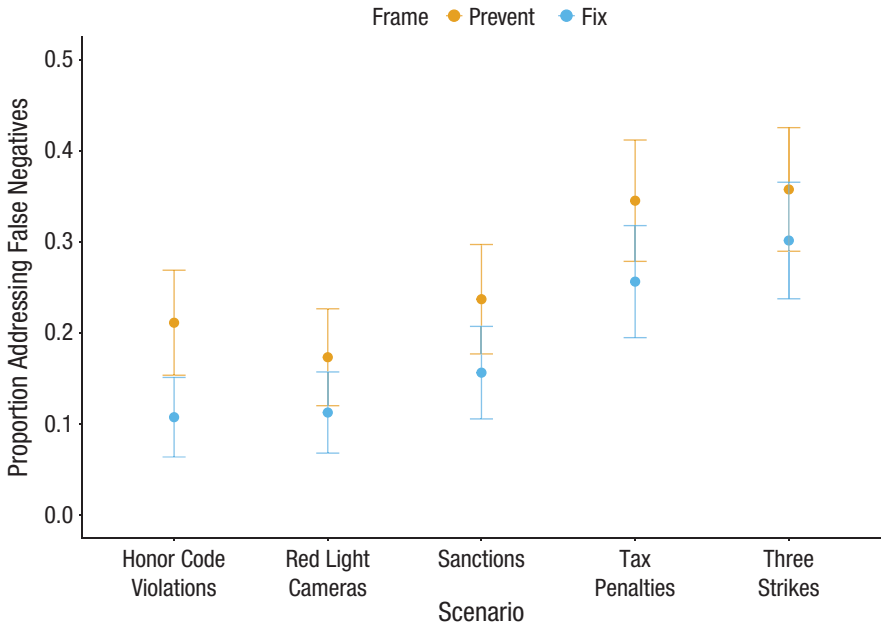

Fig. 5. Supplemental Experiment 3: real-world punishment policies. Error bars represent 95% confidence intervals.

but the future is inherently more uncertain than the past. Furthermore, we did not manipulate the reversibility or severity of errors, and often the most severe mistakes (i.e., those with the highest stakes) cannot be fixed (e.g., capital punishment). Severity might also shape the perceived or actual difficulty of addressing those mistakes. And relatedly, intuitions about base rates and preconceived notions of prevailing error rates might matter.

Another important caveat is that we exclusively recruited participants in the United States. There are well-documented issues with the generalizability of

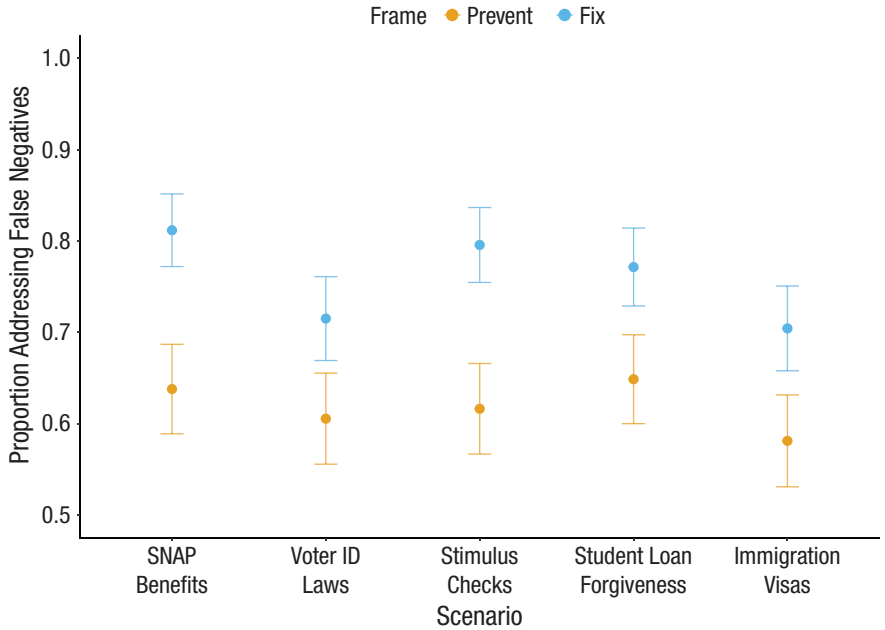

Fig. 6. Supplemental Experiment 4: real-world reward policies. Error bars represent 95% confidence intervals.

psychological phenomena beyond Western cultures (Henrich et al., 2010; Thalmayer et al., 2021). Future work might therefore explore cross-cultural differences.

Moreover, because punishments and rewards can be reframed as losses and gains, respectively, a natural consideration is the potential role of prospect theory and the related concept of reference dependence (Kahneman & Tversky, 1979). For example, participants may have hesitated to claw back false-positive rewards because doing so would have imposed losses on certain individuals. However, if participants were concerned about imposing losses, then in the context of punishments they should not have been less willing to prevent false positives than to fix them. Reference dependence thus seems limited in its ability to parsimoniously explain our results.

Overall, our framework builds upon and extends several theories. For example, research exploring the identifiable-victim effect (Schelling, 1968) has similarly implicated vividness as one potential cause (Lee & Feeley, 2016; Small, 2015; cf. Jenni & Loewenstein, 1997). Additionally, we document another critical context in which judgments about the past and future differ (e.g., Burns et al., 2012; Caruso, 2010; Cooney et al., 2016). Finally, given that the allocation of punishments and rewards can trigger concerns about fairness (Bolton & Ockenfels, 2006; Cappelen et al., 2023; Davidai & Tepper, 2023; Fehr & Schmidt, 1999; Shaddy & Shah, 2018), this work introduces another important consideration potentially shaping these perceptions.

Conclusion

Sir William Blackstone famously posited that "it is better that ten guilty persons escape than that one innocent suffer" (Blackstone, 1769). However, the present research suggests that this claim is incomplete. These preferences also depend on whether errors are considered before or after they occur, and whether they pertain to punishments or rewards. Our findings thus provide a framework for understanding seemingly inconsistent perceptions about various types of policies and offer critical insights for policymakers, managers, and marketers (among others).

Transparency

Action Editor: Mark Brandt
Editor: Patricia J. Bauer
Author Contributions

Eitan D. Rude: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Franklin Shaddy: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The authors thank UCLA Anderson's Morrison Center for Marketing and Data Analytics for financial support.

Open Practices

All preregistrations, materials, data, and code are available at: https://osf.io/69rjm/. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Eitan D. Rude https://orcid.org/0000-0001-5182-1421
Franklin Shaddy https://orcid.org/0000-0002-1153-4839

Acknowledgments

The authors thank Eugene Caruso, Ed O'Brien, Anuj K. Shah, Stephen A. Spiller, and members of the UCLA Anderson Behavioral Decision Making Lab for valuable feedback.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976241228504

References

Blackstone, W. (1769). *Commentaries on the Laws of England, Vol. II, Book IV*. Duyckinck, Long, Collins & Hannay, and Collins & Co.

Bolton, G. E., & Ockenfels, A. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, *96*(5), 1906–1911. https://www.doi.org/10.1257/aer.96.5.1906

Burns, Z. C., Caruso, E. M., & Bartels, D. M. (2012). Predicting premeditation: Future behavior is seen as more intentional than past behavior. *Journal of Experimental Psychology: General*, *141*(2), 227–232. https://doi.org/10.1037/a0024861

Cappelen, A. W., Cappelen, C., & Tungodden, B. (2023). Second-best fairness: The trade-off between false positives and false negatives. *American Economic Review*, *113*(9), 2458–2485. https://doi.org/10.1257/aer.20211015

Caruso, E. M. (2010). When the future feels worse than the past: A temporal inconsistency in moral judgment. *Journal of Experimental Psychology: General*, *139*(4), 610–624. https://doi.org/10.1037/a0020757

Caruso, E. M., Gilbert, D. T., & Wilson, T. D. (2008). A wrinkle in time: Asymmetric valuation of past and future

events. *Psychological Science*, *19*(8), 796–801. https://doi.org/10.1111/j.1467-9280.2008.02159.x

Chambers, J. R., & Davis, M. H. (2012). The role of the self in perspective-taking and empathy: Ease of self-simulation as a heuristic for inferring empathic feelings. *Social Cognition*, *30*(2), 153–180. https://doi.org/10.1521/soco.2012.30.2.153

Cooney, G., Gilbert, D. T., & Wilson, T. D. (2016). When fairness matters less than we expect. *Proceedings of the National Academy of Sciences, USA*, *113*(40), 11168–11171. https://doi.org/10.1073/pnas.1606574113

Crosby, F. J., Iyer, A., Clayton, S., & Downing, R. A. (2003). Affirmative action: Psychological data and the policy debates. *American Psychologist*, *58*(2), 93–115. https://doi.org/10.1037/0003-066X.58.2.93

Davidai, S., & Tepper, S. J. (2023). The psychology of zero-sum beliefs. *Nature Reviews Psychology*, *2*, 472–482. https://doi.org/10.1038/s44159-023-00194-9

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e1004085. https://doi.org/10.1371/journal.pmed.1004085

Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*, 235–257. https://doi.org/10.1023/A:1007740225484

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Kane, J., Van Boven, L., & McGraw, A. P. (2012). Prototypical prospection: Future events are more prototypically represented and simulated than past events. *European Journal of Social Psychology*, *42*(3), 354–362. https://doi.org/10.1002/ejsp.1866

Keller, P. A., & Block, L. G. (1997). Vividness effects: A resource-matching perspective. *Journal of Consumer Research*, *24*(3), 295–304. https://doi.org/10.1086/209511

Lee, S., & Feeley, T. H. (2016). The identifiable victim effect: A meta-analytic review. *Social Influence*, *11*(3), 199–215. https://doi.org/10.1080/15534510.2016.1216891

Lowery, B. S., Unzueta, M. M., Knowles, E. D., & Goff, P. A. (2006). Concern for the in-group and opposition to affirmative action. *Journal of Personality and Social Psychology*, *90*(6), 961–974. https://psycnet.apa.org/doi/10.1037/0022-3514.90.6.961

Munguia Gomez, D. M., & Levine, E. E. (2022). The policy–people gap: Decision-makers choose policies that favor different applicants than they select when making individual decisions. *Academy of Management Journal*, *65*(3), 842–869. https://doi.org/10.5465/amj.2020.1740

Nielsen, K., & Rehbeck, J. (2022). When choices are mistakes. *American Economic Review*, *112*(7), 2237–2268. https://doi.org/10.1257/aer.20201550

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095

Schelling, T. C. (1968). The life you save may be your own. In S. Chase (Ed.), *Problems in public expenditure analysis* (pp. 127–162). Brookings Institution.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, *84*(1), 60–79. https://doi.org/10.1037/0022-3514.84.1.60

Shaddy, F., & Shah, A. K. (2018). Deciding who gets what, fairly. *Journal of Consumer Research*, *45*(4), 833–848. https://doi.org/10.1093/jcr/ucy029

Small, D. A. (2015). On the psychology of the identifiable victim effect. In I. G. Cohen, N. Daniels, & N. Eyal (Eds.), *Identified versus statistical lives: An interdisciplinary perspective* (pp. 13–23). Oxford University Press.

Small, D. A., & Loewenstein, G. (2005). The devil you know: The effects of identifiability on punishment. *Journal of Behavioral Decision Making*, *18*(5), 311–318. https://doi.org/10.1002/bdm.507

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*, *76*(1), 116–129. https://doi.org/10.1037/amp0000622

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Van Boven, L., & Ashworth, L. (2007). Looking forward, looking back: Anticipation is more evocative than retrospection. *Journal of Experimental Psychology: General*, *136*(2), 289–300. https://doi.org/10.1037/0096-3445.136.2.289