scientific reports

Check for updates

OPEN

Diffusion of punishment in collective norm violations

Anita Keshmirian[1,2,3,4]✉, Babak Hemmatian[5], Bahador Bahrami[6,7,8], Ophelia Deroy[3,9,10] & Fiery Cushman[1]

People assign less punishment to individuals who inflict harm collectively, compared to those who do so alone. We show that this arises from judgments of diminished individual causal responsibility in the collective cases. In Experiment 1, participants (*N* = 1002) assigned less punishment to individuals involved in collective actions leading to intentional and accidental deaths, but not failed attempts, emphasizing that harmful outcomes, but not malicious intentions, were necessary and sufficient for the diffusion of punishment. Experiments 2.a compared the diffusion of punishment for harmful actions with 'victimless' purity violations (e.g., eating a dead human's flesh as a group; *N* = 752). In victimless cases, where the question of causal responsibility for harm does not arise, diffusion of collective responsibility was greatly reduced—an outcome replicated in Experiment 2.b (*N* = 479). Together, the results are consistent with discounting in causal attribution as the underlying mechanism of reduction in proposed punishment for collective harmful actions.

In 44 BCE, Roman senators plotted Julius Caesar's murder, collectively stabbing him more than 20 times at a senate meeting. Who, exactly, was to blame and to what extent? Many crimes like gang rape, collective hate crime, co-offending, and conspiracies are committed by groups. Understanding how blame and punishment are assigned in such group harms helps refine current models of moral judgment, and assess their correspondence with legal liability standards.

We dissociate two factors that might influence judgments of collective harm: intent to harm, and causal responsibility for it. Generally, people judge an actor as fully blameworthy if they intentionally cause harm[1–5]. Much research suggests that these two factors—intentionality and causal responsibility for harm—play dissociable roles in moral judgment[6–8]. They may influence the judgment of group actions in different, even contradictory, ways.

Intentionality. How are intent-based moral judgments affected by the distinction between solo and group actors? One natural possibility is that group actors are held just as responsible, given that each member of the group volitionally decides to engage in transgressive behavior. Alternatively, they may also be held less responsible on the belief that they got socially "caught up" in something they would not otherwise have done[9]. In this case, group actors would receive less blame and hence punishment than solo actors committing equivalent acts.

Causal responsibility. How will judgments of the causal responsibility of group actors compare with solo actors? Again, one possibility is that it will make no difference. Causal responsibility may be treated as categorical—one is either responsible or not[10]. Since participants in group harms are causal *contributors* to the outcome, they would be held causally responsible to the same extent as a solo actor (e.g., in felony murder[11]). For instance, in many US states (e.g., Connecticut General Statutes. tit. 53a-54c, Chapter 952; 2012), in the case of collective murder, it is argued that since the felony itself *causes* death, every participant in the felony is causally responsible for the death (Fig. 1—left panel).

Alternatively, causal responsibility may be diminished when distributed across a number of people (diffusion of punishment hypothesis; Fig. 1—right panel). This comports with several foundational ideas in the literature on causal attribution. First, "causal discounting" refers to the idea that causal attributions to one variable are

[1]Department of Psychology, Harvard University, Cambridge, USA. [2]Graduate School for Neuroscience, Ludwig-Maximilians-University of Munich, Munich, Germany. [3]Faculty of Philosophy, Ludwig-Maximilians University, Munich, Germany. [4]Munich Center for Mathematical Philosophy, Munich, Germany. [5]Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, USA. [6]Faculty of Psychology, Ludwig-Maximilians-University, Munich, Germany. [7]Department for Psychology, Royal Holloway University of London, Egham, UK. [8]Centre for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. [9]Munich Center for Neuroscience, Munich, Germany. [10]Institute of Philosophy, School of Advanced Study, University of London, London, UK. ✉email: anita.keshmirian@campus.lmu.de

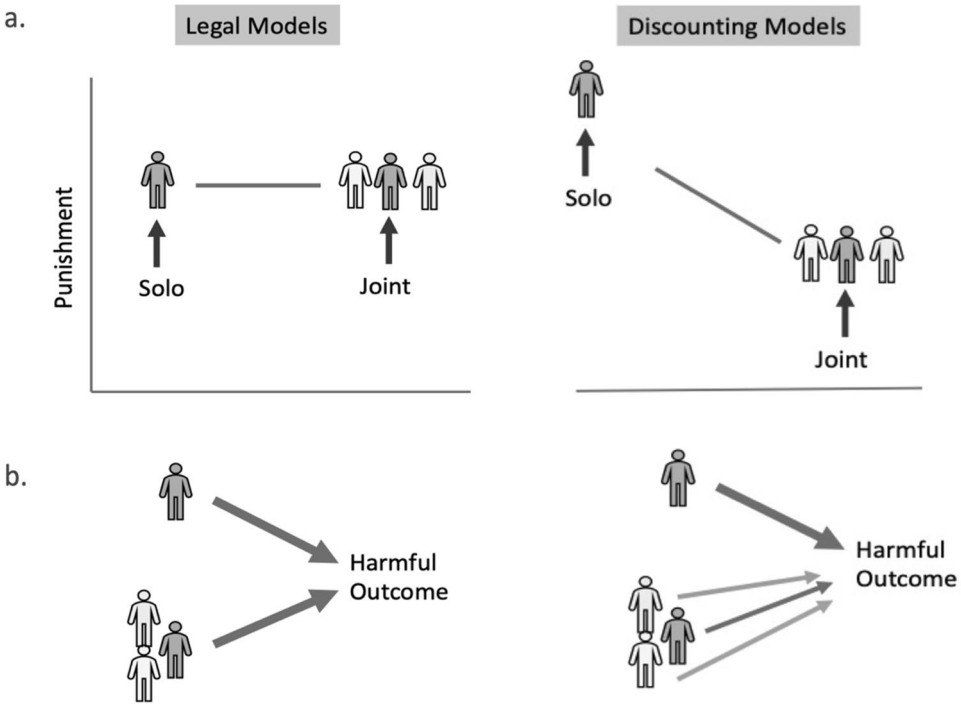

Figure 1. (**a**) Two models of punishment in solo and joint harmful actions: legal models suggest similar punishment for joint and solo acts. Discounting models predict less punishment in joint than solo harm violations (the diffusion of punishment hypothesis) (**b**) Causal links in two models of punishment. In legal models, all perpetrators in joint actions are causally responsible for the harmful outcome to the same degree as in solo actions. In the discounting models (the diffusion of punishment hypothesis), each individual in the group is less causally responsible for the outcome.

diminished as other contributing variables are introduced[12,13]. Second, "overdetermination" happens when an effect would have occurred without the contribution of any sole individual. In such cases, people are likely to perceive each individual as less causally responsible[14]. Similarly, the degree to which the individual has causal control over the outcome may be diminished in collective violations, and so causal power theory would suggest diminished attributions of causal responsibility[15]. Because punishment judgments are sensitive to attributions of causal responsibility for harm[6], these notions predict diminished punishment for group actors.

Existing research. Evidence on the punishment of groups compared to solo actors is mixed. An archival study suggested that judges give harsher sentences to lone offenders compared to group offenders, controlling for the crime[16]. However, a follow-up experiment on hypothetical robberies failed to find corroborating evidence, a result ascribed to its small sample size[16]. More recently, researchers investigated second-party punishment in fairness-based group games but found no difference between proposed punishment for lone fairness violators compared with collective ones[17]. However, second-party punishment introduces unique self-oriented emotions and motives, such as retaliation. These may bias second parties to attribute heightened intent and causal responsibility, even for groups, which could mask the diffusion of punishment. Another study about punishing cheaters showed that group violators are considered less dishonest than individual ones, but differences in judgments of deserved punishment were not statistically significant ($p = 0.08$)[18].

A key limitation of prior studies is that they cannot dissociate the potentially divergent roles of intent-based and responsibility-based processes in deserved punishment judgments. To disentangle these, in Experiment 1, we compared accidental harm (where there is no malign intent, but causal responsibility is preserved) and attempted harm (where the intent is preserved, but no harm is caused). In Experiment 2, we investigated cases of collective "harmless" purity violations, such as disrespecting the deceased, and compared them to collective harmful actions. Like attempted harms, these preserve the element of volitional action going against moral norms while eliminating any relevant question of causal responsibility.

Experiment 1

Experiment 1 tested whether a third party punishes an individual less if she inflicts a harmful outcome on a victim as part of a group, rather than acting alone. We expected less punishment assigned to individuals in a harmful joint action compared to harmful solo actions, due to a diffusion of causal responsibility for the harm. However, when the group *intended* to cause harm but no harm ensued, we did not expect to see any difference between punishment in solo versus joint actions.

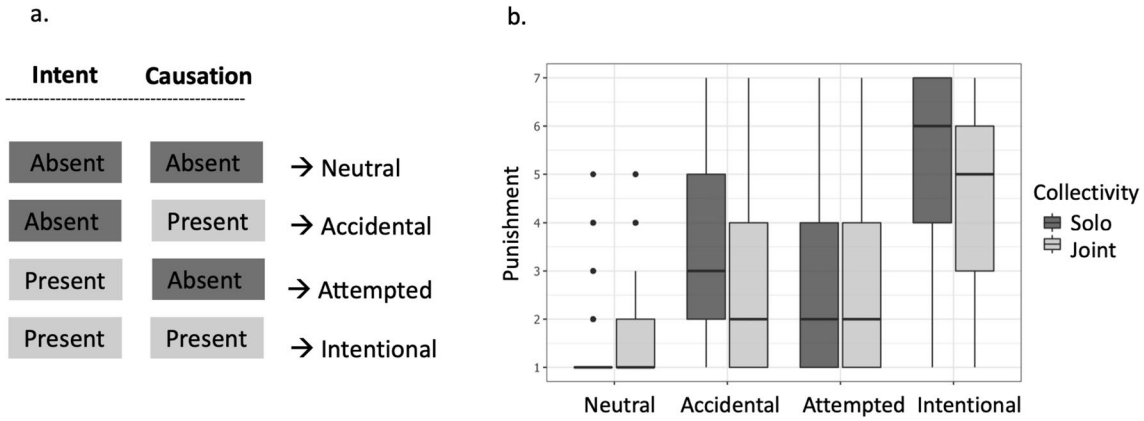

Figure 2. (**a**) The four experimental conditions as the outcome of a 2 × 2 design crossing Malicious Intent (absent vs. present) and Causation of Harm (absent vs. present). (**b**) Box-and-whisker plot of punishment ratings as a function of Collectivity (different colors) across neutral, accidental, attempted, and intentional actions (horizontal axis). The box represents the middle 50% of scores. The thick horizontal line within each box represents the median. Upper and lower whiskers show the range of scores in the highest and lowest quartiles. The dots represent outliers.

We employed a 2 × 2 × 2 design with three factors: Collectivity, Malicious Intent and Causation of Harm. Collectivity was a between-subjects factor, while Malicious Intent (henceforth called 'Intent' in short) and Causation of Harm (henceforth called 'Causation' in short) served as within-subject factors. By independently manipulating agents' Intent (absent vs. present) and Causation (absent vs. present), we can differentiate between the effects of Collectivity on Intent- and causal responsibility-based processes of moral judgment.

Methods. *Participants.* One thousand and seventy-five participants were recruited via Amazon's Mechanical Turk. Thirty-seven participants were excluded for having duplicate IDs. We used a data-driven Mahalanobis Distance measure[19] to identify non-human participants and inconsistent or inattentive responses (see Supplementary Material—Sect. 1.1). This step resulted in excluding 36 participants. We replicated the main results including those who failed the Mahalanobis exclusion criterion (see Supplementary Material—Sect. 1.1.2, Table S3). The final sample of 1002 US residents (452 males, eight choosing the "other" option) had an average age of 29.29 years ($SD = 7.46$, range: 18 to 64).

Material and procedure. Each participant was randomly assigned to one of two Collectivity conditions (joint or solo action) and read four moral scenarios in which a character committed an act either as part of a group (joint action) or alone (solo action). The dependent measure was always the deserved punishment for a given character on a 7-point scale (1 labeled as "not at all", 4 as "somewhat", and 7 as "a lot").

Intent (absent vs. present) and Causation (absent vs. present) were crossed within subjects across the four scenarios (see Fig. 2a). In neutral conditions, the agent(s) acted with no malign intention, and caused no harm. Accidental conditions involved an unintended death following the described action. In the attempted and intentional cases, the agent(s) acted with malign intent, either failing or succeeding in murdering another person. The following is an intentional, solo, harmful action scenario adapted from a previous study[8] (see Supplementary Material—Sect. 3.1 for full scenario texts):

> *Stacey* and Kate are friends and decide to go rock climbing. They are going to use new harnesses to scale a gigantic cliff.
> Kate starts to put on one of the new harnesses. The clamp on the new harness is subtly flawed, so the whole harness is incredibly unsafe to use.
> Because the clamp on the harness does not audibly click into place, *Stacey realizes* that the new harness is malfunctioning and may not be safe to use.
> *She straps Kate into the harness and asks Kate to go first.* Partway up the cliff, the harness gives way, causing Kate to fall and die.

The three sentences in italics were substituted in joint action conditions with statements about "Stacey, Anita, James, and Kate" instead, implicating the first three characters in the harm inflicted on the last-named individual.

A random pairing of stories was first created for within-subject manipulations and then counterbalanced across participants. The order of scenarios was randomized. Demographics followed the last vignette, including age, gender, political orientation (from 1 denoted as "very liberal" to 7 marked as "very conservative"), ethnicity, and education level.

Results. Figure 2b shows the results of Experiment 1. Statistical analysis was conducted using R (https://www.r-project.org/), employing generalized mixed-effects models appropriate for our design's hierarchical

structure. Since our dependent variable has a Likert scale, we employed ordinal logistic mixed-effect models using the 'ordinal' package[20].

Punishment ratings for intentional harm were significantly higher than accidental and attempted harm, and ratings in the mentioned conditions all exceeded those for the neutral condition, showing that the Intent and Causation manipulations worked as expected (see Supplementary Material—Sect. 1.2.2, Table S2).

To test the diffusion of punishment hypothesis, we modeled punishment judgments using an ordinal mixed-effects model. We included Collectivity (solo vs. group), Intent (present vs. absent), and Causation (present vs. absent) as fixed effects, along with all possible interactions. We included participant and vignette as random intercepts, along with the 'maximal' random slopes structure advocated in prior research[21]. We then performed a series of model comparisons, contrasting this full model with sparser models omitting fixed effects of interest. Model comparison favored the variant including all three factors (see Supplementary Material—Sect. 1.2.2).

The interactions between Collectivity and Causation ($b = 1.412$, $SE = 0.195$, $z = 7.245$, $p < 0.001$, two-tailed test), and between Collectivity and Intention ($b = 1.005$, $SE = 0.195$, $z = 5.159$, $p < 0.001$, two-tailed test) were significant. To better interpret the results, we computed contrasts over estimated marginal means using the 'emmeans' package in R[22]. Pairwise comparison (adjusting for multiple comparisons using the Tukey method) showed less assigned punishment for characters involved in joint actions compared to solo actions for intentional ($b = 0.574$, $SE = 0.129$, $z = 4.451$, $p < 0.001$, two-tailed test) and accidental killings ($b = 0.436$, $SE = 0.128$, $z = 3.412$, $p < 0.015$, two-tailed test), but not significant in failed murder attempts ($b = 0.029$, $SE = 0.129$, $z = 0.229$, $p = 1$, two-tailed test). Since we predicted a null effect for failed attempts, following Aczel et al.'s[23] method, a Bayesian mixed-effect analysis was performed using the 'brms' package in R[24] to confirm the pattern of results (intentional: $BF_{10} = 498.32$, $CI_{95} = [0.29, 0.69]$; accidental: $BF_{10} = 21.74$, $CI_{95} = [0.12, 0.52]$; attempted: $BF_{10} = 0.05$, $CI_{95} = [-0.17, 0.22]$; see Supplementary Material—Sect. 1.3, Table S4). In addition, to ascertain that the effect size we observe in failed attempts was small or close to zero, we performed an equivalence test using the 'TOSTER' package in R[25]. The equivalence test further confirmed that the distribution of punishment in joint vs solo attempted murders were equivalent ($z = 7.770$, $p < 0.001$, two-tailed test), indicating no significant difference between the two conditions. Unexpectedly, protagonists in neutral conditions received harsher proposed punishment for joint compared to solo actions ($b = 0.976$, $SE = 0.166$, $p < 0.001$, two-tailed test; see Fig. 2.a and Table S2 in Supplementary Material—Sect. 1.2.2). This effect was not predicted. Comparisons of specific items can be found in Supplementary Material—Sect. 1.4, Figure S2.

Discussion. We found a robust reduction in proposed punishment across instances of intended and accidental harm when perpetrators acted as part of a group rather than lone agents. The contrast between these results and previous studies[16,17] may be attributed to the more representative range of clearly and more strongly harmful Causations (i.e., death) represented in our materials. That no diffusion of punishment was observed for attempted harm suggests that diffusion of punishment depends on the discounting principle involved in *causal* attribution of harmful *outcomes* rather than intentions.

Experiment 2

Not all acts deemed immoral involve causing harm. 'Victimless' purity violations are condemned on the basis of a moral norm violation rather than harmful causation[25]. They are judged based on perpetrators' impact on themselves rather than victims[27–29], and elicit disgust only when people judge moral character rather than outcomes[30]. If the diffusion of punishment results from a discounting principle in causal attribution, it would only apply to actions that cause harmful outcomes. Therefore, we expected it to be weaker for judgments of purity violations.

We test this prediction in Experiment 2 by directly comparing diffusion of punishment in scenarios involving harm and purity violations. Unlike most previous studies, instead of assuming that harm scenarios induce a sense of harmfulness alone and the purity vignettes only a sense of disgust, we asked participants to rate how harmful or gross they found the protagonist(s)' action in all scenarios, examining their evaluation of the causation more directly.

Experiment 2.a. *Methods.* Participants. A target sample size was predetermined using a Monte Carlo simulation following guidelines provided by DeBruine and Barr[31] (see Supplementary Material Sect. 2.5.1). The final sample consisted of 752 US and UK residents (331 females; three others, age: $M = 28.03$ years, $SD = 6.63$, range: 18 to 60) recruited through Prolific Academic (https://www.prolific.co/) and compensated for their time. Twenty-six participants were excluded for having the same Prolific IDs as those from a pilot study. To increase precision, we used data-driven methods (as in Experiment 1) to exclude inattentive responders. Another 39 were excluded for failing attention checks—they assigned 0 to 49 (on a 100-point scale) blame to a person who "destroys the entire planet" ($n = 18$), and 51 to 100 for someone who "gives money to a charitable organization" ($n = 21$).

Materials and procedure. Experiment 2a compared judgments for less grave *harm* than in Experiment 1 (e.g., intentionally breaking someone's leg) with *purity* violations (e.g., masturbating over a grave). Collectivity was manipulated as before. The moral Domain (harm vs. purity) was manipulated within subjects. Hence, each subject responded to four scenarios, presented to her in two blocks (for harm and purity, respectively). The scenarios were randomly chosen from a battery of 8 items, counterbalanced across participants. The order of blocks and the items within each block were randomized and counterbalanced. Four items were adapted from a previous study on harm[8], while four items were original scenarios representing purity violations (some inspired by a previous study[32]; see Supplementary Material—Sect. 3.2 for the full text of scenarios). The items were matched for severity of joint vs. solo action in a pilot study.

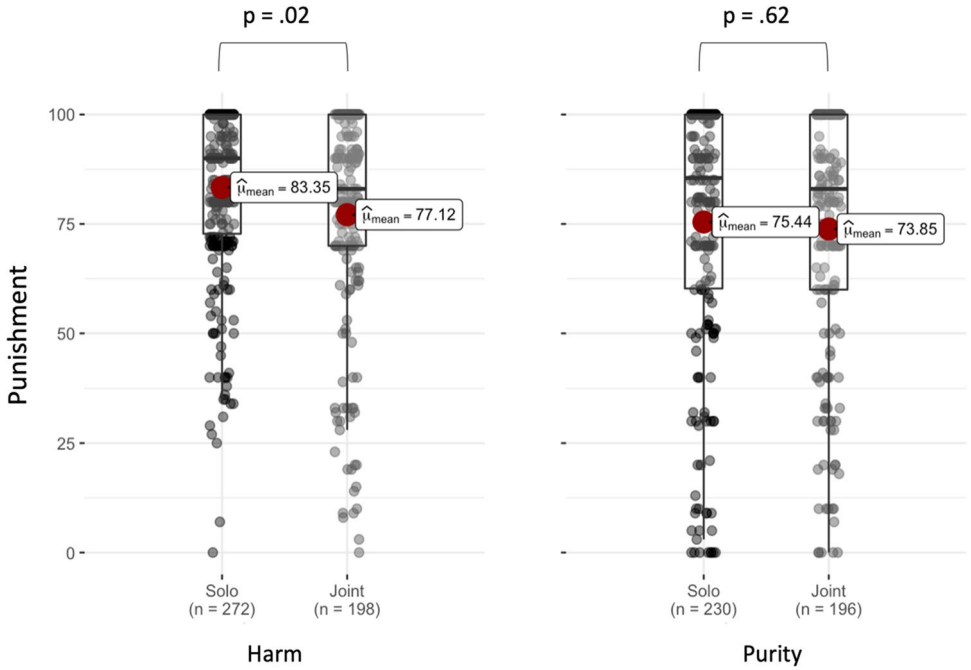

Figure 3. Harmfulness (left) and Grossness (right) ratings are matched across Joint and Solo actions but significantly different across Domains: Harmfulness is higher in Harm (left), and grossness is higher in Purity (right) scenarios. Graph conventions are the same as Fig. 2.

For instance, a purity violation in the solo condition would read:

> Dan's favorite singer has died and has been buried in a nearby cemetery. He always had wild fantasies about the singer, and one night, he forms the following plan: He enters the cemetery late at night and goes to masturbate over the singer's grave, making sure he cannot be seen. After that, he ensures that the grave is clean and exactly as it was before and leaves.

The same scenario in the joint condition would introduce Dan, Ray, and Carl as friends who collectively committed the act.

To allow more precise judgments by providing more response options, we measured proposed punishment on a 100-point Likert scale. Zero to 50 was labeled as mild and 50 to 100 as severe punishment. Perceived Harmfulness and Grossness were measured on similar 100-point scales. Judgments of blameworthiness (hereafter "Blame") were also gathered in Experiment 2.a to rule out an alternative explanation for the diffusion of punishment: if the practical difficulty of punishing multiple violators compared with a single actor is behind the diffusion, Blame judgments that are more removed from such practicalities would not show the effect. The results were in line with our non-pragmatic interpretation of diffusion, despite some differences with punishment ratings (see Supplementary Material—Sect. 2.3).

Like Experiment 1, we employed generalized ordinal mixed-effects models appropriate for our design's hierarchical structure. Since our dependent variable was on a 100-point scale, we employed linear mixed-effect models through the 'LME4' package in R[33]**.** Other aspects of the design were identical to Experiment 1.

Results. We designed the scenarios with the goal of minimizing perceived harm in Purity conditions and perceived grossness in Harm conditions. However, no scenario garnered an average perceived Harmfulness rating of close to zero. Therefore, we first tested whether perceptions of harm and purity violation show the expected results, regardless of researcher-assigned labels for scenario Domains. Punishment remained the key outcome measure throughout this analysis.

We modeled Harmfulness and Grossness ratings using a linear mixed-effects model with Collectivity (solo vs. group) and Domain (purity vs. harm) as fixed effects, along with their possible interactions. Participant and vignette were included as random intercepts, along with 'maximal' random slopes[21]. Model comparisons contrasting this full model with sparser models ensued. The full model showed the best performance. Contrasts over estimated marginal means were calculated using the 'emmeans' package[22].

As expected, across joint and solo norm violations, significant effects of Domain were found for Harmfulness ($b = 40.032$, $SE = 4.686$, $z = 8.542$, $df = 6.372$, $p < 0.001$, two-tailed test) and Grossness ($b = 41.454$, $SE = 2.430$, $z = 17.059$, $df = 7.366$, $p < 0.001$, two-tailed test), indicating that participants perceived Harm scenarios as significantly more harmful (and less gross) than Purity scenarios (see Fig. 3). No significant effect of Collectivity was found for Harmfulness ($b = 0.531$, $SE = 1.497$, $z = 0.354$, $df = 1324.678$, $p = 0.722$; two-tailed test; $BF_{10} = 0.754$, $CI_{95} = [-1.68, 1.44]$) or Grossness ratings ($b = 0.573$, $SE = 1.727$, $z = 0.087$, $df = 1150.680$, $p = 0.704$; two-tailed test; $BF_{10} = 0.76$, $CI_{95} = [-1.57, 2.26]$), whether using linear mixed-effects analysis[33] or its Bayesian

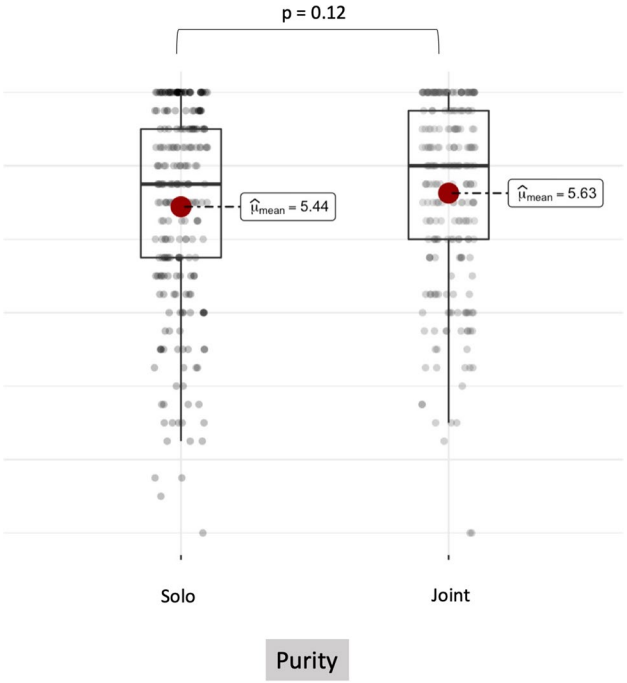

Figure 4. In the first blocks, participants punished individuals for Solo actions (dark grey) more than Joint actions (light grey) in Harm scenarios (left) but not in Purity conditions (right).

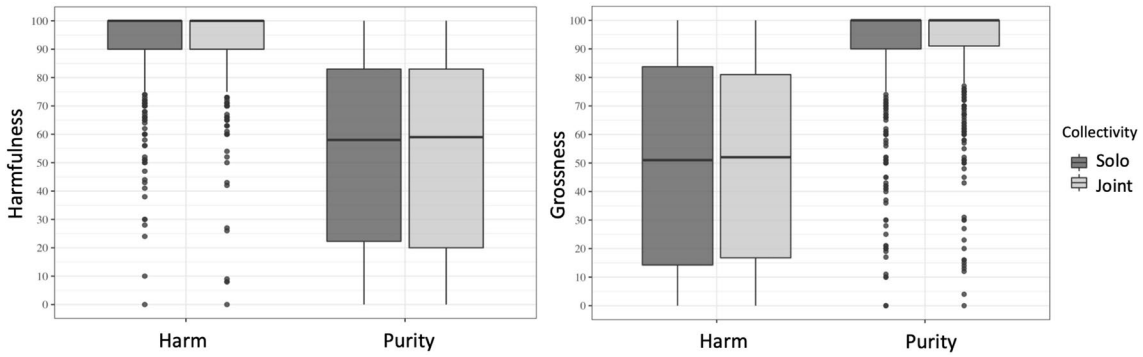

Figure 5. No difference in punishment judgments was observed between Solo and Joint Purity violations in Experiment 2.b.

counterpart[24]. Because the interaction between Domain and Collectivity was not significant in Harm ($b = 1.270$, $SE = 1.649$, $z = 0.77$, $df = 2053.909$, $p = 0.442$ two-tailed test) or in Purity ($b = 1.408$, $SE = 1.696$, $z = 0.830$, $df = 2052.974$, $p = 0.406$, two-tailed test), we did not further calculate a main effect of Domain. In addition to establishing the adequacy of our item construction, these results support the Moral Foundations account of disparate moral domains for harm and purity[34].

We then investigated the interaction between Collectivity (Solo vs. Joint) and Domain (Purity vs. Harm) in a linear mixed-effect model predicting punishment ratings, including participant and vignette as random intercepts along with 'maximal' random slopes[21](see Supplementary Material—Sect. 2.1.1). There was a main effect of Collectivity ($b = 3.364$, $SE = 1.519$, $z = 2.214$, $df = 1139.669$, $p = 0.027$, two-tailed test) but the predicted interaction with domain was not significant ($b = 0.181$, $SE = 1.480$, $z = 0.122$, $df = 2053.622$, $p = 0.902$, two-tailed test). Given the within-subjects blocked design, we investigated possible spillover effects in which participant responses to the first block (of, e.g., harm cases) would affect their responses in the second block (of, e.g., purity scenarios). This might have the effect of deflating diffusion of punishment in harm cases and inflating it in purity cases. To confirm the dissociation of Domains, we dissected the first blocks into Harm and Purity datasets. An exploratory mixed effect analysis (not preregistered) with a model similar to the above –but only including responses to the first blocks—showed the predicted significant diffusion of punishment in Harm ($b = 6.005$, $SE = 2.648$, $z = 2.266$, $df = 232.942$, $p = 0.024$, two-tailed test), but not Purity blocks ($b = 1.742$, $SE = 3.503$, $z = 0.497$, $df = 209.431$, $p = 0.619$, two-tailed test) (see Fig. 4; for more details see Supplementary Material—Sect. 2.1.1). To further confirm that the effect size we observed in Purity cases was small or close to zero, we performed an

equivalence test[25]. As predicted, the equivalence test was significant ($z = 4.58$, $p < 0.001$, two-tailed test), showing that punishment distributions in Joint and Solo violations of Purity are equivalent in the first blocks.

We used this dataset to test the Collectivity by Domain interaction in a linear mixed-effect model where Domain (Harm vs Purity) was introduced as a between-subject factor. Using the combined Harm and Purity dataset, rather than dissecting the first blocks into two independent sets, allows us to maintain greater statistical power. The predicted interaction with Domain was not significant in the combined first-block dataset *(*$b = 4.251$, $SE = 4.347$, $z = 0.978$, $p = 0.328$, $df = 442.213$, two-tailed test*)*. A post-hoc power simulation based on the observed parameters showed that our sample size was inadequate to reliably detect a plausible effect size for the interaction (see Supplementary Material—Sect. 2.5). To ensure adequate power and confirm that carryover effects are responsible for the observed diffusion in the Purity domain, we conducted a pre-registered replication with only purity violations in a between-subjects design.

Experiment 2.b. *Methods Exp 2.b.* Participants. A target sample size of 500 was predetermined using a Monte Carlo simulation via the 'SIMR' package in R[35]. We determined the sample size with 93.00% (91.24, 94.50) power to detect the main effect of Collectivity in the Harm domain with the same parameters obtained from the main regression model of punishment in Experiment 2.a. We recruited 526 US and UK residents to ensure that, after exclusions, the sample size will be close to the target. Participants were recruited through Prolific Academic (https://www.prolific.co/) and compensated for their time. Thirty-nine entries were excluded for having Prolific IDs that duplicated those from a pilot study. Attention checks and their results were as in Exp 2.a. Eight participants were excluded for failing attention checks. The final sample consisted of 479 (312 females; five others, age: $M = 27.53$ years, $SD = 6.57$, range: 18 to 62).

Materials and procedure. The number of protagonists was manipulated as in Experiment 2.a, but only one moral Domain (Purity) was provided to the subjects. Each participant responded to four fully randomized scenarios, all from the Purity domain. The scenarios were identical to Experiment 2.a. We measured deserved punishment after each scenario on a 7-point Likert scale similar to Experiment 1 since the 100-point scale had no impact on the results in Experiment 2.a.

Results. A linear mixed-effect analysis was performed with Collectivity (Joint vs. Solo) as a fixed factor, and participants and vignettes as random factors. Pairwise comparison indicated that judgments were similar in Joint and Solo purity violations ($b = 0.175$, $SE = 0.119$, $z = 1.464$, $df = 476.997$, $p = 0.143$; two-tailed test), which was confirmed by Bayesian mixed-effects analysis ($BF_{10} = 0.377$, $CI_{95} = [-0.429, 0.041]$) (see Fig. 5 and Supplementary Material—Sect. 2.2). To further confirm that the effect size we observed in Purity cases was small or close to zero, we performed an equivalence test. Using TOSTER package in R[25]. As predicted, the equivalence test was significant ($z = 8.738$, $p < 0.001$, two-tailed test), showing that punishment distributions in Joint and Solo violations of purity are equivalent.

Discussion. In Experiment 2.a, we found a stronger diffusion of punishment in the Harm domain, along with evidence that any diffusion in Purity scenarios may be due to carryover effects. Confirming this interpretation, our preregistered Experiment 2.b found no diffusion of punishment for actions deemed impure but harmless, despite ample power. This suggests that punishment is diffused only when the collective action contains a causal link to a harmful outcome, and is absent for victimless moral violations.

General discussion

Though group immoral actions are commonly performed, punitive reactions to them are rarely studied. Research on solo actions shows that punishment depends on two general processes: judgments of causal responsibility and the intent to harm[6]. Drawing on two complementary and well-established paradigms for dissociating causal and intent-based processes, we studied how punishment judgments respond to collective moral violations.

In Experiment 1, a reduction of punishment in group harmful acts was attributable to the causal process of moral judgment—a diffusion of causal responsibility. This is consistent with discounting theories which argue that assigning punishment follows from a causal attribution of harmful outcomes, whereby having more than one sufficient cause results in lower responsibility assigned to each cause[4,12,14,36–38]. In contrast, we found no reduction in punishment attributable to the *intent-based* process of moral judgment.

Two different methods provided convergent evidence for the dissociation between causal and intent-based contributions to judgments of group crimes. First, we found that accidental harm-doers (who bear causal responsibility for harm without intent) were punished less when part of a group compared to solo actors. Yet attempted harms (acting with harmful intent but bearing no causal responsibility for harmful outcomes) were punished identically across solo and collective contexts. Second, we found that having an identifiable harmed victim was necessary for the diffusion: victimless purity violations were punished equivalently across Collectivity conditions.

The diffusion of punishment observed for harmful outcomes can explain how individuals use group membership to minimize senses of regret and responsibility[17], and protect themselves from the costs of moral violations like punishment[39]. Seeking 'safety in numbers' by acting as part of groups, each perpetrator may expect mitigated punishment and blame. The diffusion may therefore promote collective moral norm violation[40–42].

Our findings also bear on theories of moral judgment. First, they support the dissociation of causal and mental-state processes in moral judgment[6–8,32]. Second, they support disparate judgment processes for harmful versus "victimless" moral violations[26–30,32]. Third, they reinforce the idea that punishment often involves a "backward-looking", retributive focus on responsibility, rather than a "forwards-looking" focus on rehabilitation, incapacitation, or deterrence (which, we presume, would generally favor treating solo and group actors

equivalently). Punishers' future-oriented self-serving motives and their evolutionary roots need further investigation as alternative sources for the diffusion of punishment. For instance, punishing joint violators may produce more enemies for the punisher, reducing the motivation for a severe response.

Whether the diffusion of punishment and our causal explanation for it extends to other moral domains (e.g., fairness[43]) is a topic for future research. It is also possible that Purity violations induce a diffusion of punishment as well, but one that is masked by a corresponding increase in the perceived severity of joint purity offenses. Examining this possibility is a task for future research. Another interesting extension is whether different causal *structures* produce different effects on judgments. Our vignettes were intentionally ambiguous about causal chains and whether multiple agents overdetermined the harmful outcomes. Contrasting diffusion in conjunctive moral norm violation (when collaboration is *necessary*) with disjunctive ones (when one individual would *suffice)* is informative, since attributions of responsibility are generally higher in the former class[4,12–14,36,44].

Our findings highlight a divergence between legal theories of justice and laypeople's perceptions of apt punishment when harm is inflicted collectively, shedding light on the cognitive underpinnings of collective atrocities in the hopes of a more moral future. Whether and how the discrepancy can be addressed may have implications for society at large.

Ethical approval. Experiment 1 has been reviewed for compliance with ethical research standards and approved by the Harvard University Ethics Committee under the umbrella protocol (IRB14-2016). Protocols of experiment 2.a and 2.b were approved by London School of Advanced Study Research Ethics Committee (approval ref. SASREC_1819_313A). All methods were carried out in accordance with relevant guidelines and regulations of Harvard University (Experiment 1) and School of Advanced Study, University of London (Experiment 2.a and 2.b) ethics committees.

Informed consent. In all experiments reported in the manuscript, informed consent was obtained from all participants.

Statement of relevance

When judging a crime committed by a group, each criminal's liability must be determined individually. How do ordinary people judge the punishment suitable for collective harmful actions? We show that an individual harm-doer's punishment is typically, but not always, reduced in collective actions. The exceptions point to an underlying cognitive mechanism. We observe diffusion of punishment when punishment depends on an assessment of causal responsibility for harm, such as in cases of intentional or accidental harm. Diffusion of punishment is absent in cases where there is no causal responsibility for harm, e.g., in unsuccessful attempts or victimless actions that are deemed immoral. These findings refine current cognitive accounts of punishment, and have implications in forensic settings, allowing us to contrast the structure of legal liability with the moral intuitions of ordinary people.

Data availability

De-identified data for all experiments, along with a codebook and materials, are openly available at https://osf.io/m3f47/. The preregistration for experiment 2.a and 2.b can be accessed at https://osf.io/hjnxm and https://osf.io/xw39e respectively.

Received: 1 April 2022; Accepted: 24 August 2022
Published online: 12 September 2022

References

1. Guglielmo, S., Monroe, A. E. & Malle, B. F. At the heart of morality lies folk psychology. *Inquiry* **52**(5), 449–466. https://doi.org/10.1080/00201740903302600 (2009).
2. Malle, B. F., Guglielmo, S. & Monroe, A. E. A theory of blame. *Psychol. Inq.* **25**(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340 (2014).
3. Malle, B. F. & Knobe, J. The folk concept of intentionality. *J. Exp. Soc. Psychol.* **33**(2), 101–121. https://doi.org/10.1006/jesp.1996.1314 (1997).
4. Shaver, K. G. *The Attribution of Blame: Causality, Responsibility, and Blameworthiness* (Springer-Verlag, 1985).
5. Shultz, T. R. & Wright, K. Concepts of negligence and intention in the assignment of moral responsibility. *Can. J. Behav. Sci.* **17**(2), 97–108. https://doi.org/10.1037/h0080138 (1985).
6. Cushman, F. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* **108**(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006 (2008).
7. Young, L., Cushman, F., Hauser, M. & Saxe, R. The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci. U.S.A.* **104**(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104 (2007).
8. Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A. & Saxe, R. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc. Natl. Acad. Sci. U.S.A.* **107**(15), 6753–6758. https://doi.org/10.1073/pnas.0914826107 (2010).
9. Malle, B. F., Moses, L. J. & Baldwin, D. A. Intentions and intentionality: Foundations of social cognition. *The MIT Press* https://doi.org/10.7551/mitpress/3838.001.0001 (2001).
10. Moore, M. S. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics* (Oxford University Press, 2010).
11. Binder, G., Weisberg, R. & Fissell, B. M. Capital punishment of unintentional felony murder. Social Science Research Network. *Notre Dame Law Rev.* **92**, 1141–1214 (2017).
12. Kelley, H. H. The processes of causal attribution. *Am. Psychol.* **28**(2), 107–128. https://doi.org/10.1037/h0034225 (1973).
13. Morris, M. W. & Larrick, R. P. When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychol. Rev.* **102**(2), 331–355. https://doi.org/10.1037/0033-295X.102.2.331 (1995).
14. Lagnado, D., Gerstenberg, T. & Zultan, R. Causal responsibility and counterfactuals. *Cogn. Sci.* **37**(6), 1036–1073. https://doi.org/10.1111/cogs.12054 (2013).

15. Cheng, P. W. From covariation to causation: A causal power theory. *Psychol. Rev.* **104**(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367 (1997).
16. Feldman, R. S. & Rosen, F. P. Diffusion of responsibility in crime, punishment, and other adversity. *Law Hum Behav.* **2**(4), 313–322. https://doi.org/10.1007/BF01038984 (1978).
17. El Zein, M., Seikus, C., De-Wit, L. & Bahrami, B. Punishing the individual or the group for norm violation. *Wellcome Open Res.* **4**, 139 (2020).
18. Vainapel, S., Weisel, O., Zultan, R. & Shalvi, S. Group moral discount: Diffusing blame when judging group members. *J. Behav. Decis. Mak.* **32**(2), 212–228. https://doi.org/10.1002/bdm.2106 (2019).
19. Dupuis, M., Meier, E. & Cuneo, F. Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behav. Res. Methods* **51**(5), 2228–2237. https://doi.org/10.3758/s13428-018-1103-y (2019).
20. Christensen, R. H. B. Ordinal - Regression Models for Ordinal Data. R package version 2019.3–9 (2019). http://www.cran.r-project.org/package=ordinal/.
21. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001 (2013).
22. Lenth, R.V. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.2 (2022). https://CRAN.R-project.org/package=emmeans
23. Aczel, B. *et al.* Quantifying support for the null hypothesis in psychology: An empirical investigation. *Adv. Methods Pract. Psychol. Sci.* https://doi.org/10.1177/2515245918773742 (2018).
24. Bürkner, P. Bayesian item response modeling in R with brms and Stan. *J. Stat. Softw.* **100**(5), 1–54. https://doi.org/10.18637/jss.v100.i05 (2021).
25. Lakens, D. Equivalence testing with TOSTER. *APS Observer*, 30 (2017).
26. McHugh, C., McGann, M., Igou, E. R. & Kinsella, E. L. Searching for moral dumbfounding: Identifying measurable indicators of moral dumbfounding. *Collab. Psychol.* **3**(1), 23 (2017).
27. Chakroff, A., Dungan, J. & Young, L. Harming ourselves and defiling others: What determines a moral domain?. *PLoS ONE* **8**(9), e74434. https://doi.org/10.1371/journal.pone.0074434 (2013).
28. Chakroff, A., Russell, P. S., Piazza, J. & Young, L. From impure to harmful: Asymmetric expectations about immoral agents. *J. Exp. Soc. Psychol.* **69**, 201–209. https://doi.org/10.1016/j.jesp.2016.08.001 (2017).
29. Dungan, J. A., Chakroff, A. & Young, L. The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PLoS ONE* **12**(3), e0173405. https://doi.org/10.1371/journal.pone.0173405 (2017).
30. Giner-Sorolla, R. & Chapman, H. A. Beyond purity: Moral disgust toward bad character. *Psychol. Sci.* **28**(1), 80–91. https://doi.org/10.1177/0956797616673193 (2017).
31. DeBruine, L. M. & Barr, D. J. Understanding mixed-Effects models through data simulation. *Adv. Methods Prac. Psychol. Sci.* https://doi.org/10.1177/2515245920965119 (2021).
32. Rottman, J. & Young, L. Specks of dirt and tons of pain: Dosage distinguishes impurity from harm. *Psychol. Sci.* **30**(8), 1151–1160. https://doi.org/10.1177/0956797619855382 (2019).
33. Bates, D., Mächler, M., Bolker, B. & Walker, S. fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**(1), 1–48. https://doi.org/10.18637/jss.v067.i01 (2015).
34. Graham, J. *et al.* Moral foundations theory: The pragmatic validity of moral pluralism. *Adv. Exp. Soc. Psychol.* **47**, 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4 (2012).
35. Green, P. & MacLeod, C. J. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* **7**(4), 493–498. https://doi.org/10.1111/2041-210X.12504 (2016).
36. Gerstenberg, T. & Lagnado, D. Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition* **115**(1), 166–171. https://doi.org/10.1016/j.cognition.2009.12.011 (2010).
37. Halpern, J. *Actual Causality* (The MIT Press, 2016).
38. Kelley, H. H. Causal schemata and the attribution process. In *Attribution Perceiving the Causes of Behaviour* (eds Jones, E. E. *et al.*) (Lawrence Erlbaum Associates, Inc, 1987).
39. El Zein, M., Bahrami, B. & Hertwig, R. Shared responsibility in collective decisions. *Nat. Hum. Behav.* **3**(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4 (2019).
40. Bandura, A., Underwood, B. & Fromson, M. E. Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *J. Res. Pers.* **9**(4), 253–269. https://doi.org/10.1016/0092-6566(75)90001-X (1975).
41. Darley, J. M. & Latane, B. Bystander intervention in emergencies: Diffusion of responsibility. *J. Pers. Soc. Psychol.* **8**(4), 377–383. https://doi.org/10.1037/h0025589 (1968).
42. Latane, B. *et al.* Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.* **37**(6), 822–832 (1979).
43. Graham, J. *et al.* Mapping the moral domain. *J. Pers. Soc. Psychol.* **101**(2), 366–385. https://doi.org/10.1037/a0021847 (2011).
44. Zultan, R., Gerstenberg, T. & Lagnado, D. Finding fault: Causality and counterfactuals in group attributions. *Cognition* **125**(3), 429–440. https://doi.org/10.1016/J.COGNITION.2012.07.014 (2012).

Acknowledgements

We acknowledge Courtney McQuade, Justin Sulik, Stephan Sellmaier, Ulrike Hahn, Stephan Hartmann, Steven Sloman, Neurophilosophy Group at Graduate School Of Neuroscience, Morteza Dehghani, Alexander Soutschek, Alexandre Zenon and Javad Hatami.

Author contributions

A.K. generated the study concept. Data collection was performed by A.K. for experiments 1 and 2.a and 2.b and A.K. and B.H. for experiment 2.a. A.K. performed interpretation and visualization of both experiments. A. Keshmirian prepared materials for Experiment 1. A.K. and O.D. prepared materials for Experiment 2 while B.B. and B.H. provided critical revisions. Experiment 1 was performed under the supervision of F.C. and Experiments 2a and 2b under the guidance of F.C., B.B., and O.D., A.K. analyzed the data of both experiments and drafted the original manuscript, while B.H., B.B., O.D., and F.C. provided critical revisions. F.C., O.D., and B.B. provided funding.

Funding

This research was funded by a grant from SEED (to FC), by the NOMIS foundation (to OD). An overview of Experiment 2 was presented at Oxford University at The Ninth International Symposium on "Biology of Decision Making" in May 2019. BB was supported by the Humboldt Foundation, the NOMIS Foundation, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant

agreement No. 819040—acronym: rid-O). A.K was supported by DFG grant HA 3000/21-1, the priority program SPP 1999: Robust Argumentation Machines (RATIO). B.H was affiliated with the Department of Cognitive, Linguistic and Psychological Sciences at Brown University at the time of the study and is currently affiliated with the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-19156-x.

Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.