Attribution of Responsibility Between Agents in a Causal Chain of Events

Vanessa Cheung* (vanessa.cheung.14@ucl.ac.uk)

Mengxuan Qiao* (helen.qiao.18@ucl.ac.uk)

David Lagnado (d.lagnado@ucl.ac.uk)
Department of Experimental Psychology, University College London,
26 Bedford Way, WC1H 0AP, London, UK
* These authors contributed equally.

Abstract

In this paper, we explored the attribution of causal responsibility in a causal chain of events, where an agent A instructs an intermediate agent B to execute some harmful action which leads to a bad outcome. In Study 1, participants judged B to be more causally responsible, more blameworthy, and more deserving of punishment than A. In Study 2, we explored the effect of proximity on judgments of the two agents by adding a third, subsequent contributing cause, such that B's action no longer directly caused the final outcome. Participants judged both agents A and B to be less causally responsible and deserving of punishment (but not less blameworthy) when they were less proximal to the outcome, and there were no differences in judgments between the two agents. In Study 3, we varied whether each of the two agents (A and B) intended for the final outcome to occur. We find an interaction between role and intent, where participants only mitigated judgments for A when A did not intend for the outcome to occur – regardless of B's intent. We discuss possible explanations for our findings and its implications for moral and legal decision-making.

Keywords: moral judgment; moral decision-making; causal models; blame attribution; causal responsibility; causal chain

Introduction

In February 2022, Jennifer Faith, who had instructed her boyfriend Darrin Lopez to kill her husband, pleaded guilty to a murder-for-hire charge – an offense that carries a potential death sentence in the state of Texas[1]. In July 2023, Darrin Lopez was convicted of murder[2]. While there are a number of considerations for how the two defendants are sentenced in court, many of them specific to the situation, we are interested in understanding folk perceptions of causal responsibility in such cases where an individual commits harm 'by proxy'. Do laypeople consider Faith and Lopez to be equally responsible and blameworthy for the victim's death? Do they think that the two deserve equal punishment? These are questions related to how people attribute responsibility to multiple agents when they have contributed differently to an outcome.

One key difficulty in evaluating causality is that different factors can interact and combine in various ways to cause an outcome. Causal judgments are often graded, and people attribute responsibility differently depending on how causal factors interact with each other (Gerstenberg & Lagnado, 2010). The causal structures of scenarios can take many forms (DeScioli & Kurzban, 2013), and in cases that feature multiple agents, it is often difficult to determine which agent is more responsible for the outcome and by how much.

There are three causal structures where two agents A and B may be responsible for a single outcome: (1) both agents independently contribute to cause an outcome; (2) the actions of A cause the actions of B which directly cause the outcome, and (3) the group, of which A and B are both part of, jointly causes the outcome (Kaiserman, 2021). While many studies have explored how people attribute responsibility in (1) (e.g., Gerstenberg & Lagnado, 2010), fewer have explored situations such as (2) where B's actions depend on A in a causal chain (see Figure 1).

This is a case of joint causation, where both agents are necessary for the outcome. While A did not directly cause the outcome, it would not have occurred but for their instigation; on the other hand, while B physically caused the outcome to occur, they would not have done so in the first place had it not been for A. Consider the case described earlier: do people consider Faith more responsible as none of the subsequent events would have happened *but for* her actions? Or do they consider Lopez more responsible because he *directly and physically* caused the victim's death?

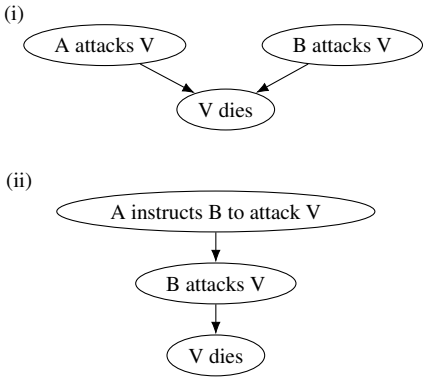

Figure 1: The two models illustrate the causal relationship between A, B, and V. The top model (i) is a common-effect model, where A and B jointly and independently contribute to the outcome of V's death. The bottom model (ii) is a causal chain model, where A causes an intermediate event B contributing the outcome of V's death.

[1]https://www.justice.gov/usao-ndtx/pr/jennifer-faith-pleads-guilty-murder-hire-husband-s-death

[2]https://www.nbcnews.com/news/us-news/darrin-lopez-jennifer-faith-love-triangle-murder-army-veteran-rcna117945

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Joint Responsibility in a Causal Chain

In a causal chain, while prior events do not directly cause the outcome to occur, transitivity of causation implies that if some event A causes B, and B causes an outcome (O), it follows that A had caused O (Hilton et al., 2010). However, people do not judge these events to be equally causal: they tend to attribute more causality to the proximal cause than the distal cause (e.g., Lagnado & Channon, 2008; Reuter et al., 2014). These causal evaluations also allow people to make moral judgments (Sloman et al., 2009; Waldmann et al., 2017) – they also often judge the proximal cause to be less morally permissible (Engelmann & Waldmann, 2022).

Situations with multiple *agents* are invariably more complex, especially in moral or legal contexts where a large variety of factors come into play. Legal theorists have proposed that if there is a 'free and deliberate' human action in the causal chain, it is more likely to be selected as an explanation for an outcome than a physical cause (e.g., natural causes) (Hart & Honoré, 1985). Imagine a scenario where two agents, A and B, jointly cause an outcome: A deliberately throws a lighted cigarette into a shrub which catches fire, and the fire is exacerbated by B pouring petrol on the flames. Hart and Honoré (1985) argue that people are more likely to attribute causality to the proximal cause (B) since it is a human action that intervened in the causal chain, regardless of whether the distal cause (A) is a human action or a physical cause. Conversely, if the proximal cause was a physical cause (e.g., instead of B pouring petrol, the wind had exacerbated the flames), then people are more likely to attribute causality to A's actions. Other studies have supported this theory, finding that people prioritise voluntary actions as explanations over physical causes (Hilton et al., 2010; McClure et al., 2007).

However, note that each agent acted independently in the above example: B did not act in concert with A, nor did A's actions bring about B's actions (McClure et al., 2007). As such, this chain of events may be better described as a temporal chain rather than a direct causal chain. Other studies investigating causal chains have also tended to focus on similar situations where B acts unknowingly or under some situational constraint brought about by A (Fincham & Shultz, 1981; Phillips & Shaw, 2015). To our knowledge, no research has focused on causal and moral judgments in situations where A and B act in concert with each other, but take on different roles – such as that of instigator and executor of a criminal act. Therefore, we aim to investigate how people make causal and moral judgments on agents taking two different roles (i.e., instigator and executor) in a causal chain.

Causal Chains in the Law

As far as legal causation is concerned, a defendant's act must be an operative and substantial cause of the outcome. However, the 'chain' of causal connection between the person who caused the original act and the subsequent outcome can be broken by an intervening cause (*novus actus interveniens*). The interruption of a causal chain occurs when a third party intervenes by an action that is free, deliberate, and informed. The presence of such an intervening cause (e.g., an act committed by a second, proximal agent) relieves the person who started the chain of events (i.e., the distal agent) from responsibility for causing the outcome. However, this is not always applied consistently; for instance, it can be unclear whether an act is truly voluntary or informed (Firkins, 2023).

One crucial criterion for determining whether the intervening action breaks the causal chain lies in its reasonable foreseeability (Law, 2015). Naturally, in a straightforward 'hired gun' murder, the victim's death is a foreseeable outcome to the instigator. But imagine a scenario where A only instructs B to *attack* V, not necessarily intending V's death (as in Figure 1). Would the causal chain break if B develops their own motive of wanting V dead, which leads to an action (and outcome) that A could not have foreseen? In other words, would laypeople mitigate A's responsibility when B's action goes beyond what A had intended? Another interesting question arises as to whether a subsequent act by a third party (e.g., the victim themselves) may break the causal chain between the defendants' act and the outcome. In the law, such interruptions to the causal chain can relieve responsibility of the distal agent(s) – but do laypeople make judgments of responsibility consistent with these assumptions? We aim to answer these questions in this paper, and build on existing work on how folk and legal ascriptions of causation are similar and where they differ (e.g., Güver & Kneer, 2023; Knobe & Shapiro, 2021).

Overview of Studies

Across three studies, we investigated how people attribute causal responsibility, blame, and punishment to two agents in a causal chain. We presented participants with a vignette based on a real murder case, in which two defendants, an instigator (A) and an executor (B) of the criminal act contribute jointly to kill a victim (V). In this scenario, A does not directly harm V but instructs B to do so.

In Study 1, we investigate the effect of an agent's role (in the causal chain) on causal and moral judgments using a short scenario based on a legal case. Further, we explore the role of proximity and intent using variations of this scenario. We vary the proximity of the two agents' contributions to the outcome (i.e., how much B directly caused the outcome; Study 2) and whether each agent intended for the outcome to occur (Study 3), to explore how these factors influence laypeople's causal and moral judgments. All materials, data, and analyzes are available at the OSF repository (https://osf.io/bg5xz/).

Study 1

First, we investigated whether people would assign the same or different levels of causal responsibility, blame, and punishment depending on an agent's role in a causal chain of events.

Method

To achieve 80% power for detecting a medium effect size at an alpha-level of 0.05, 36 participants were required. We re-

cruited 90 UK participants from Prolific (M_{age} = 38.5, SD_{age} = 13.2, 45 male, 45 female; paid £0.70).

Design The study used a within-subjects design. We measured participants' judgments of causal responsibility, blame, and punishment for the two agents (A and B). They were shown two scales for each measure on a scale of 0-100, one for giving a response for A, and the other for B.

Materials and Procedure We presented an original vignette in an online experiment on Qualtrics. This described a legal case, loosely based on *R. v Rook* (Court of Appeal, 1993), where A ('Adam') instructed a 'hired gun', B ('Ben'), to attack V ('Veronica'), which led to her death. Both parties admitted to their actions to reduce ambiguity about credibility. The vignette is shown below:

> *Adam Smith and Ben Parker are on trial for the death of Veronica Brown. In the trial, it was stated that Ben caused severe injuries to Veronica in her home on Thursday 10 November 2022. According to the forensic report, Veronica died soon from the injuries, and her body was found in the early hours of the following day.*
>
> *Evidence showed that Adam used an online platform to initiate contact with Ben prior to the victim's death. On Monday 7 November (three days before the victim's death), the pair had met in person. During their conversation, Adam provided specific instructions for the attack, and told Ben to cover it up so it would look like an accident.*
>
> *At the trial, Ben admitted to attacking Veronica, but stated that he would not have done so if Adam had not instructed him to. Adam insisted that despite his involvement, the killing was ultimately executed by Ben and that Ben's actions were the main cause of Veronica's death.*

Participants then gave their responses to the questions *'To what extent was each defendant causally responsible / blameworthy / deserving of punishment for the victim's death?'* Responses to the first two questions was made on a scale of 0 (Not at all) to 100 (Completely), and for punishment on a scale of 0 (No punishment) and 100 (Maximum punishment [i.e., life imprisonment]). The order of the agents' names was displayed in counterbalanced order between participants.

They answered these questions twice, with two sets of instructions presented in counterbalanced order. In one version of the questions, participants could freely give ratings to each defendant (i.e., free allocation). In the other version, participants were told that for each measure, their responses for the two defendants could only total 100 (i.e., fixed sum). The purpose was to control for potential effects driven by different instructions (Kaiserman, 2021).

Results

First, we adjusted the free allocation ratings to be comparable with the fixed sum ratings by re-scaling the former to show the proportion of scores given to each agent. For each participant, we then averaged across the fixed sum and re-scaled free allocation ratings to obtain one rating for each agent. We conducted paired t-tests to compare mean judgments for agents A and B. These tests were conducted separately for each judgment measure (causal responsibility, blame, and punishment).

Overall, participants were more severe in their moral judgments for B compared to A. Participants judged B ($M = 58.4$) to be more causally responsible than A ($M = 41.6$), $t(89) = 5.88$, $p < .001$, $d = 0.62$. They also judged B ($M = 54.2$) to be more blameworthy than A ($M = 45.9$), $t(89) = 4.09$, $p < .001$, $d = 0.43$. Further, they judged B ($M = 54.4$) to deserve more severe punishment than A ($M = 45.6$), $t(89) = 4.67$, $p < .001$, $d = 0.49$. Figure 2 visualizes the differences in judgments between A and B.

Figure 2: Mean proportion of causal responsibility, blame, and punishment ratings distributed to each agent in Study 1. Error bars indicate 95% CI.

We find similar results in an analysis of only free allocation responses, showing that participants intuitively make similar judgments without being prompted to compare the two agents. Participants judged B ($M = 89.9$) to be more causally responsible than A ($M = 75.0$), $t(89) = 5.35$, $p < .001$, $d = 0.56$. They also judged B ($M = 91.0$) to be more blameworthy than A ($M = 83.4$), $t(89) = 3.79$, $p < .001$, $d = 0.40$. Further, they judged B ($M = 87.7$) to deserve more severe punishment than A ($M = 79.9$), $t(89) = 3.99$, $p < .001$, $d = 0.42$.

Study 2

Study 1 revealed that participants judged the proximal agent (the executor, B) to be more causally responsible, blameworthy, and deserving of more severe punishment than the distal agent (the instigator, A). We further explore this proximity effect in Study 2, where we extend the causal chain by adding a third, more proximal cause (V's own behaviour) that more directly contributes to the final outcome (V's death). We investigate people's judgments of the two agents when B's actions are no longer most proximal to the final outcome.

Method

Participants To achieve 80% power for detecting a medium effect size at an alpha-level of 0.05, 86 participants were required. We recruited 99 UK participants on Prolific (M_{age} = 37.4, SD_{age} = 13.8, 49 male, 50 female; paid £0.60).

Design and Materials This study used a 2 (between-subjects; proximity) × 2 (within-subjects; role) design. Participants were randomised into two groups and saw two different versions of the same vignette. In the 'high proximity' condition, participants saw the same vignette as in Study 1 where B directly causes the final outcome. In the 'low proximity' condition, they saw a similar vignette except V survives the attack. V's leg is permanently damaged by the attack, and she later dies from falling down a flight of stairs due to this injury. We piloted the condition with the new vignette prior to the main experiment ($N = 50$; see OSF repository).

Procedure We used the same measures and procedure as Study 1, except participants also rated the causal responsibility and blameworthiness of the victim (not included in the main analysis). They allocated all ratings freely. Participants also answered the question *'How foreseeable do you think V's death was to each person?'* on a scale of 0-100.

Results

Main Analysis To test the effect of proximity on moral judgments for agents A and B, we conducted 2 (proximity) × 2 (role) mixed ANOVA for judgments of causal responsibility, blameworthiness, and punishment separately using *afex* (Singmann et al., 2022), and pairwise comparisons using *emmeans* (Lenth, 2022) with Tukey adjusted p-values.

Overall, participants judged both agents to be more causally responsible, $F(1,97) = 4.09, p = .046$, and deserved more severe punishment, $F(1,97) = 9.07, p = .003$, when they directly caused the outcome (i.e., high proximity). There was no effect of proximity on judgments of blame, $F(1,97) = 3.23, p = .076$. Participants in the high proximity condition also judged the outcome to be more foreseeable to both agents, $F(1,97) = 21.45, p < .001$.

Within the high proximity condition, we replicated our findings on causal responsibility from Study 1: participants judged the two agents differently. Here, they judged B (M = 91.4, 95% CI[86.0, 96.8]) to be more responsible than A (M = 84.8, 95% CI[78.7, 90.9]), $t(97) = 3.02, p = .017$. In the low proximity condition, we found no difference in judgments between the two agents, $t(97) = 1.23, p = .609$. Figure 3 visualizes the interaction between proximity and role, but this was not significant, $F(1,97) = 1.57, p = .214$.

Contrary to Study 1, we found no significant differences between blame judgments for agents in the high proximity condition, $t(97) = 1.01, p = .741$, and also no differences between agents in the low proximity condition, $t(97) = 1.55, p = .412$. The interaction between proximity and role was not significant, $F(1,97) = 3.30, p = .072$.

We again found no significant differences between judgments of punishment for the two agents in the high proximity condition, $t(97) = 2.08, p = .166$, and the low proximity condition, $t(97) = 0.79, p = .858$. The interaction between proximity and role was not significant, $F(1,97) = 0.82, p = .368$.

Cross-Study Analysis We conducted an additional analysis using the data aggregated from Studies 1 and 2 ($N = 140$) for the high proximity condition. Overall, B was more causally responsible, $t(139) = 5.92, p < .001, d = 0.50$, blameworthy, $t(139) = 3.84, p < .001, d = 0.33$, and deserving of punishment, $t(139) = 4.33, p < .001, d = 0.37$ than A. For those measures that did not replicate in Study 2, we added 'study' as a variable and found no significant interaction between the samples of the two studies and role in judgments of blame, $F(1,138) = 3.06, p = .082$, and punishment, $F(1,138) = 2.97, p = .087$). It is therefore plausible that we did not replicate these findings due to lower power for these measures in the smaller sample (79% and 83% respectively).

Discussion

Overall, the presence of a subsequent contributing cause mitigated judgments for both A and B: *both agents* were more responsible and deserving of punishment when they *more directly* caused the outcome. This suggests that participants may have attributed more responsibility to B in Study 1 because he was the most proximal cause. This is consistent with previous findings that actions are more morally permissible in a longer chain of events (Engelmann & Waldmann, 2022); however, considering that the difference in causal responsibility across proximity conditions was only just statistically significant, we plan to replicate this finding in a future study.

When proximity is reduced, participants appear to view both A and B's actions as equally indirect, as they are equally causally responsible and deserving of punishment for the outcome. V's actions might thus be seen as an intervening cause that mitigates judgments for both agents. However, this 'intervention' does not fully break the causal chain between the agents and the outcome. Neither agent is 'relieved' from the responsibility of causing the outcome: both are still highly responsible (mean ratings are all > 75).

Further, both agents were highly blameworthy regardless of proximity to the outcome. This is supported by previous findings that people tend to blame others for having bad intentions or desires even if they did not directly cause a harmful outcome (Young & Tsoi, 2013; Young & Saxe, 2011). By extending the causal chain, we also find an effect of foreseeability on moral judgments: both A and B are less responsible when participants think they did not foresee the outcome. This is consistent with previous research (Engelmann & Waldmann, 2022; Fincham & Shultz, 1981; Lagnado & Channon, 2008).

One question that follows is whether agents would also be considered less responsible if they did not *intend* for the final outcome to occur – as unintended outcomes are usually less foreseeable than intended ones (Kovacevic et al., 2024). In our vignette, the agents' intentions were ambiguous: we only mentioned that A instructed B to *attack* V, and it is therefore unclear whether either agent had intended for V to die. It is possible that inferences about both agents' intentions would influence moral judgments. Variance in these assumptions may have also contributed to why we could not replicate re-

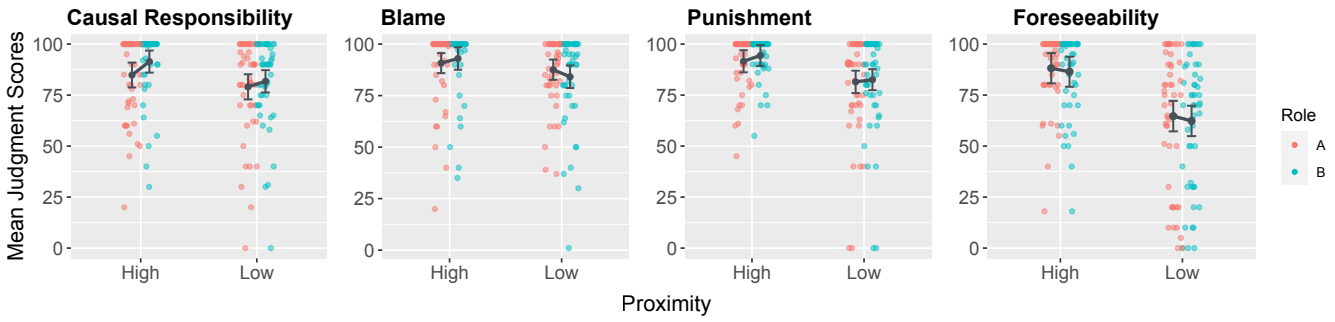

Figure 3: Mean moral judgment ratings for each agent at different levels of proximity in Study 2.

sults from Study 1. We therefore aimed to further explore this in Study 3.

Study 3

In Study 3, we explore the interaction between intent and role by directly manipulating the intent of each agent. We showed participants the same causal chain scenario as Study 1 – A instructs B to attack V, which directly leads to her death. However, we varied whether each defendant intended for the final outcome to occur (i.e., the victim's death). We explored (1) whether people would judge the agents differently depending on their intent, and (2) whether the causal chain between A and the outcome would weaken when B acts in a way that is not intended by A (and therefore less foreseeable).

Method

Participants To achieve 80% power for detecting a medium effect size at an alpha-level of 0.05, 136 participants were required. We recruited 145 UK participants from Prolific (paid £2.90 for the 19-minute study). We excluded nine participants (one for failing the attention check and eight for not answering all questions). The final sample size was $N = 136$ (M_{age} = 38.7, SD_{age} = 13.2, 65 male, 71 female).

Design and Materials This study used a 2 (between-subjects; A's intent) × 2 (between-subjects; B's intent) × 2 (within-subjects; role) design. Note that 'intent' refers to intending the final outcome (i.e., killing V), not just intending the act (i.e., attacking V). We showed a modified version of the Study 1 vignette with four different variations of intent:

- Condition 1: Both A and B intended to kill V *(A instructed B to kill V, B executed the killing)*;
- Condition 2: Both A and B did not intend to kill V *(A instructed B to 'rough up' V, B intended to 'rough up' but accidentally killed V)*;
- Condition 3: A intended to kill V but B did not *(A instructed B to kill V, B intended to 'rough up' but accidentally killed V)*;
- Condition 4: A did not intend to kill but B intended to kill V *(A instructed B to 'rough up' V, B intended to kill V and executed the killing)*.

In all versions, the outcome is the same: V dies as a result of B's actions, which were brought about by A. For the within-subjects measure (role), participants gave moral judgments for both defendants as in the previous studies.

Procedure This study also investigated how participants represented the events in causal models, but we omit these results for brevity. In the study, participants read the vignette, drew a causal model of events, and gave moral judgments.

Results

To test the effects of role and intent, we conducted a mixed 2 (A's intent) × 2 (B's intent) × 2 (role) ANOVA for each measure (causal responsibility, blame, punishment) separately.

We replicated the results of Study 1 showing that the agents' role had a significant main effect on judgments. Participants judged B to be more causally responsible, $F(1,132) = 28.87$, $p < .001$, more blameworthy, $F(1,132) = 9.56$, $p = .002$, and deserving of more severe punishment than A , $F(1,132) = 17.03$, $p < .001$.

We found that B's intent did not influence moral judgments between the two agents (causal responsibility: $F(132) = 0.61$, $p = .436$; blame: $F(132) = 0.33$, $p = .568$; punishment: $F(132) = 1.88$, $p = .173$). However, as shown in Figure 4, we found a significant interaction between A's intent and role for judgments of causal responsibility, $F(1,132) = 13.10$, $p < .001$, blameworthiness, $F(1,132) = 7.98$, $p = .005$, and how much they deserved to be punished, $F(1,132) = 22.09$, $p < .001$. We conducted pairwise comparisons between judgments for A and B at different levels of A's intent.

For causal responsibility judgments, participants judged A ($M = 80.3$, 95% CI[75.1, 85.5]) to be less responsible than B ($M = 95.4$, 95% CI[92.1, 98.7]) when A had low intent, $t(132) = 6.27$, $p < .001$. Conversely, when A had high intent, there was no significant difference in judgments between A ($M = 88.9$, 95% CI[83.8, 94.0]) and B ($M = 91.8$, 95% CI[88.7, 95.0]), $t(132) = 1.26$, $p = .592$.

For blameworthiness judgments, participants judged A ($M = 85.2$, 95% CI[80.8, 89.6]) to be significantly less blameworthy than B ($M = 94.0$, 95% CI[90.7, 97.4]) when A had low intent, $t(132) = 4.13$, $p < .001$. However, when A had high intent, judgments of blame did not significantly

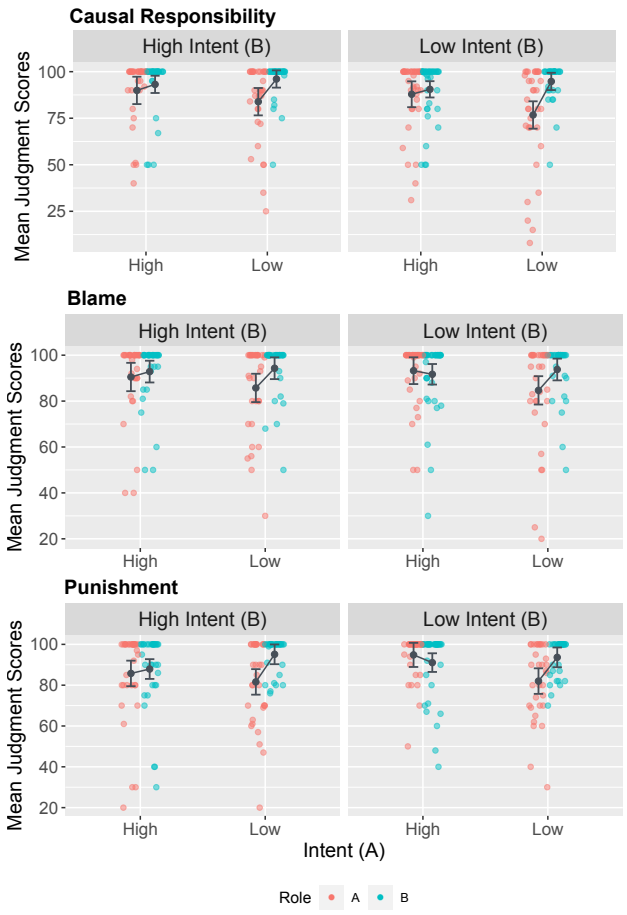

Figure 4: Mean moral judgment ratings for each agent at different levels of intent in Study 3.

differ between A ($M = 91.9$, 95% CI[87.6,96.1]) and B ($M = 92.3$, 95% CI[89.0,95.5]), $t(132) = 0.19$, $p = .998$.

For punishment judgments, participants judged A ($M = 81.8$, 95% CI[77.4,86.2]) to be less deserving of punishment than B ($M = 94.4$, 95% CI[91.0,97.8]) when A had low intent, $t(132) = 6.16$, $p < .001$. However, when A had high intent, judgments again did not significantly differ between A ($M = 90.3$, 95% CI[86.0,94.6]) and B ($M = 89.5$, 95% CI[86.2,92.8]), $t(132) = 0.41$, $p = .977$.

Discussion

Our findings replicate those of Study 1, showing that the executor (B) was generally more responsible, blameworthy, and deserving of punishment than the instigator (A). We also found an interaction between role and intent, such that people only mitigate judgments for the *instigator* when he did not intend the outcome – regardless of the executor's intentions.

These results are consistent with *novus actus interveniens* in criminal law, where a voluntary intervening cause only breaks the chain of causation when it was not reasonably foreseeable to the distal agent. When A had intended for V to die, then B's actions were reasonably foreseeable to him. Conversely, when A only intended for B to harm V and not kill her, then one could argue that B's subsequent killing of V (regardless of whether it was intentional) was less foreseeable, making A less responsible for V's death. The increased causal responsibility for A's intentional actions is also consistent with the legal realist view that intentional wrongfulness lengthens the reach of legal cause (Knobe & Shapiro, 2021).

General Discussion

In three studies, we explored how laypeople attribute causal responsibility in a causal chain where an agent (A) instructs an intermediate agent (B) to execute a harmful action resulting in the death of a victim (V). Study 1 showed that participants tended to judge B more causally responsible, more blameworthy, and more deserving of severe punishment than A. Study 2 revealed that when the proximity for A and B decreases, attributions of causal responsibility and punishment reduced for both agents, highlighting the role of proximity in these attributions. Further, Study 3 showed that irrespective of B's intent, participants mitigated their judgments for A when the latter did not intent for the outcome to occur.

Overall, we show that laypeople's moral judgments are somewhat consistent with what the law presumes: relatively unforeseeable, more proximal causes weaken, but do not break, the chain of causation. In Study 1 we found higher causal responsibility for the proximal over the distal cause, which aligned with past research (Lagnado & Channon, 2008). Studies 2 and 3 further illuminate this pattern: when a third, intervening cause contributes to the outcome, the responsibility of the two initial agents diminishes; similarly, once B's behaviour surpasses A's intent and foresight, the causal chain between A and V is weakened by B's intervention. In both cases, proximity mitigates responsibility for the outcome. This is also generally consistent with prior work on how physical indirectness (i.e., intermediation) mitigates moral judgments (Cushman et al., 2006; Engelmann & Waldmann, 2022; Paharia et al., 2009). Therefore, in situations where the resulting harm is beyond what was intended, people might intuitively be more lenient when delivering verdicts for defendants who conduct harm through an intermediate agent compared to defendants who cause harm directly.

To further explore people's causal reasoning, in future studies we will investigate responsibility attribution in scenarios with more complex causal structures and with other types of actions (e.g., negligence from omissions). In addition, we will analyze participants' qualitative responses about their reasoning process to identify common themes they consider when making moral judgments. One limitation of the current research is that all vignettes featured a severe outcome (death), which may explain the ceiling effect in some judgments. To confirm the generalizability of our findings, we will conduct studies using a range of different vignettes with varying contexts and outcomes. These future studies can further elucidate folk attribution of causal responsibility.

Acknowledgments

This work received support from the Institute for Humane Studies under grant number IHS017859 to V.C. We thank Kyoungbin Ellen Yoon for contributing to the materials and data collection for Study 3, Nichola Raihani for feedback on materials, and Teneille Brown, Yvonne McDermott Rees, Maximilian Maier, and anonymous reviewers for comments on a previous version of this manuscript.

References

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological bulletin*, *139*(2), 477.

Engelmann, N., & Waldmann, M. R. (2022). How causal structure, causal strength, and foreseeability affect moral judgments. *Cognition*, *226*, 105167.

Fincham, F. D., & Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology*, *20*(2), 113–120.

Firkins, G. (2023). Rethinking causation in english criminal law. *The Journal of Criminal Law*, *87*(1), 18–38.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Güver, L., & Kneer, M. (2023). Causation and the silly norm effect. *Advances in Experimental Philosophy of Law*, 133–168.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. OUP Oxford.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*(3), 383–400.

Kaiserman, A. (2021). Responsibility and the 'pie fallacy'. *Philosophical Studies*, 1–20.

Knobe, J., & Shapiro, S. (2021). Proximate cause explained: An essay in experimental jurisprudence. *The University of Chicago Law Review*, *88*(1), 165–236.

Kovacevic, K. M., Bonalumi, F., & Heintz, C. (2024). The importance of epistemic intentions in ascription of responsibility. *Scientific Reports*, *14*(1), 1183.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770. https://doi.org/10.1016/j.cognition.2008.06.009

Law, J. (2015). A dictionary of law.

Lenth, R. V. (2022). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.8.1-1]. https://CRAN.R-project.org/package=emmeans

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, *37*(5), 879–901.

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, *109*(2), 134–141.

Phillips, J., & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science*, *39*(6), 1320–1347.

Reuter, K., Kirfel, L., Van Riel, R., & Barlassina, L. (2014). The good, the bad, and the timely: How temporal order and moral judgment influence causal selection. *Frontiers in Psychology*, *5*, 1336.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). *Afex: Analysis of factorial experiments* [R package version 1.1-1]. https://CRAN.R-project.org/package=afex

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. *Psychology of Learning and Motivation*, *50*, 1–26.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. *Moral Inferences*, 37–55.

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass*, *7*(8), 585–604. https://doi.org/10.1111/spc3.12044

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214. https://doi.org/10.1016/j.cognition.2011.04.005