The dynamics of responsibility judgment: Joint role of dependence and transference causal explanations

Reasoning about underlying causal relations drives responsibility judgments: agents are held responsible for the outcomes they cause through their behaviors. Two main causal reasoning approaches exist: dependence theories emphasize statistical relations between causes and effects, while transference theories emphasize mechanical transmission of energy. Recently, pluralistic or hybrid models, combining both approaches, have emerged as promising psychological frameworks. In this paper, we focus on causal reasoning as involved in third-party judgements of responsibility and on related judgments of intention and control. In particular, we used a novel visual paradigm to investigate the combined effects of two well-known causal manipulations, namely omission and pre-emption, on these evaluations. Our findings support the view that people apply a pluralistic causal reasoning when evaluating individual responsibility for negative outcomes. In particular, we observed diminished responsibility when dependence, transference, or both fail, compared to when these mechanisms are upheld. Responsibility judgement involves a cognitive hybrid of multiple aspects of causal reasoning. However, important differences exist at the interindividual level, with most people weighting transference more than dependence.

Keywords: Omission, pre-emption, responsibility, causal reasoning, dependence, transference

1. Introduction

Imagine that you hear that Peter has physically assaulted your friend Ted out of jealousy for his achievements. It might come natural to you to feel resentful towards Peter. Once you overcome your immediate feelings, you might pause to examine the situation thoroughly. In particular, you may find yourself determining whether Peter is truly responsible for harming Ted and is then an appropriate target for your reproach. Did Peter intend to harm Ted or should the episode be recast as an accident? Did Peter manage to hurt Ted or did Ted come through the whole thing unscathed? Was Peter alone or did someone else contribute to hurting Ted? Did Peter understand that hurting

Ted was morally wrong and that Ted would have suffered? Many have suggested that holding someone responsible implies: (1) making these multi-faceted evaluations about whether an agent is a fair target of reproach and – if so – eventually (2) responding, e.g., by blaming the agent (Scanlon, 2008; Shoemaker, 2011). As P. F. Strawson did not fail to acknowledge, this net of reactive attitudes, including resentment and blame, is the cognitive scaffolding of everyday interpersonal relationships (1962). In this study, we aimed to investigate a subset of key factors that influence people's responsibility evaluations, and the related judgments of intention and control, in third-party scenarios.

Existing empirical research acknowledges that causal reasoning plays a particularly critical role in responsibility evaluations for negative outcomes (Cushman & Young, 2011; Pizarro et al., 2003; Shaver, 1985; Waldmann, 2017). It is widely accepted that, to decide whether Peter is responsible for hurting Ted, we must first determine whether his behavior played any causal role in Ted's being hurt. If Ted injured himself while practicing alone in the kitchen, Peter would not be a fair target of reproach. Whereas it is clear that causal reasoning plays a role, the key questions concern what causal variables and types of causal reasoning affect responsibility evaluation. In this study we shed further light on the impact of two known causal variables on responsibility, intention, and control judgments about repeatedly presented scenarios. These are the omission effect (action vs. omission) and the pre-emption effect (preventable outcome vs. non-preventable outcome), organized in a 2×2 experimental setup.

In doing so, we make no claim of being exhaustive. Several other, causal and non-causal, factors may contribute to modulating our everyday evaluations. However, these two variables will prove useful in arbitrating between two discussed types of causal reasoning, i.e., dependence and transference theories, making distinguishable

predictions about people's causal representations. In the rest of this section, we first provide a sketch of transference and dependence, and then discuss how pre-emption and omission are problematic for them.

1.1 Behavioral variables and modes of causal reasoning

Across different metaphysical frameworks and psychological explanations, causation relies on a relation between the antecedent cause and the consequent effect. The most thoroughly developed accounts of causation are transference and dependence theories. Here, we are not concerned with what causal framework best represents causation in general. Rather, we focus on what type of causal reasoning is at work when people evaluate third-party responsibility, intention, and control. Dependence and transference have often been treated as rival views, ensuing different predictions about people's causal representations. Dependence can hardly accommodate pre-emption, whereas transference has trouble with omissions. In recent years, hybrid or pluralistic accounts, combining dependence and transference, have emerged in philosophy (Hall, 2004; Hitchcock, 2003) and psychology (Lombrozo, 2010; Walsh & Sloman, 2005). More precisely, while pluralistic views predict that different causal representations are independently used in different contexts, hybrid views suggest that they can be applied within the same causal reasoning and context (Waldmann & Mayrhofer, 2016). With this paper we aim to contribute to this debate, by testing the validity of an account where dependence and transference are integrated within people's understanding of analogous types of causal sequences.

1.1.1 Dependence theories and pre-emption cases

Dependence theories view causation as a relation between two events such that the second event depends upon the first. Sophisticated versions of dependence include

probabilistic (Hitchcock, 1993) and interventionist (Woodward, 2008) accounts. In their basic form, dependence is grounded in counterfactual explanations of the relation between the cause and the effect. Lewis (1973a) suggested that one event (B) causally depends on another (A) if and only if (i) both A and B happened, and (ii) had (A) not happened, (B) would not have happened either. Psychological research has repeatedly identified counterfactual dependence as a key type of causal explanation, elucidating how people model causal relations in the environment (Kahneman & Tversky, 1982; Kulakova et al., 2017; Gilbert et al., 2015; McCormack et al., 2011). Assuming that responsibility relies on causal evaluations, holding Peter responsible thus depends upon determining whether Ted would have been hurt or not had Peter behaved differently.

The original counterfactual account is subject to conceptual problems. In particular, pre-emption cases, where dependence fails but the relation between the two events is seemingly still causal, are the bugbear of counterfactual views. Pre-emption refers to situations where, had the outcome not been caused by its actual cause, it would have been caused anyway by another event (Collins, 2000; Lewis, 1973b), e.g., had Peter not hurt Ted, Ed would have hurt Ted. It is then false that, had Peter behaved differently, the outcome would not have happened. Nonetheless, we still have the intuition that Peter's behavior has caused Ted's being hurt (see Gerstenberg et al., 2021; Stephan & Waldmann, 2018).

Pre-emption is extensively investigated in philosophy, since it is problematic for dependence accounts (Menzies, 1989; Schaffer, 2003; Yablo, 2002). Halpern and Pearl (2005) supplement Lewis' simple dependence by considering possible (i.e., non-actual) situations in which counterfactual dependence still holds. For example, there is a possible situation in which Ed does not act, and Peter acts. In this sense, Ted's being hurt counterfactually depends on Peter's action. The problem can thus be fixed at a

conceptual level.

However, given that dependence is central to causal reasoning (and thus to responsibility attributions), the failure of simple dependence in cases of pre-emption (Goldvarg & Johnson-Laird, 2001; Stephan & Waldmann, 2018) is likely to affect the extent to which one is held responsible: if an event is judged as less causal when simple dependence fails, one would expect that responsibility also diminishes compared to when simple dependence is upheld. Returning to our example, if Ted is going to get hurt anyway, Peter is likely to be judged as less responsible for hurting him compared to when Ted's being hurt necessarily depends on Peter's behavior. Along these lines, Chockler and Halpern (2004) define the degree of responsibility of an event *e* for causing an outcome as a function of the minimal number of changes that must be made to the actual situation to make it the case that the outcome counterfactually depends on the event in question (i.e., so that, had *e* not occurred, the outcome would not have occurred either).

Dependence and its failures are explored in the psychological literature as well. In a pioneering study, Wells and Gavanski showed that, to judge the causal role of prior events, people mentally simulate counterfactuals that can undo the current outcome. Their Mental Simulation Model (MSM) suggests that causal judgments arise from a comparison between actual events and a mental simulation of what would have happened in a relevant counterfactual world (Wells & Gavanski, 1989; see also the relevant work by Gerstenberg and colleagues, who developed a similar model called the *Counterfactual Simulation Model* (Gerstenberg et al., 2012; Gerstenberg et al., 2014; Gerstenberg et al., 2015; Gerstenberg et al., 2021; Lagnado et al., 2013)). The MSM predicts that the more a simulated default alternative (i.e., the most likely one) has the potential to alter the outcome, the more the actual causal event is considered causally

relevant. If the default alternative cannot undo the outcome (e.g., as due to pre-emption), the causal relevance of the actual event is judged as low. On a similar line, Petrocelli and colleagues suggested that the impact of counterfactual thinking on causal reasoning varies depending on the perceived degree to which a particular event could have been different (2011).

To sum up, assuming that people consistently apply dependence criteria to responsibility evaluations, pre-emption is expected to diminish the agent's responsibility. Within our design, we test for this factor referring to it as outcome preventability (preventable outcome vs. non-preventable outcome). Clearly, pre-emption is not the only situation in which dependence fails. Notably, in overdetermination cases, the outcome is jointly caused by the activity of multiple agents, each of which would have been sufficient to cause the effect, causing the effect simultaneously. For example, Chang (2009) investigates people's causal intuitions in an overdetermination scenario where two trains simultaneously hit a house of cards (see also Henne et al., 2019). An advantage of pre-emption, as compared to symmetric overdetermination, is that it depends on a clear temporal distinction between events. This helps to avoid ambiguities about the extent to which single agents contributed to the final outcome's being produced in the actual world (Beyer et al., 2017; Gerstenberg & Lagnado, 2010). Pre-emption also differs from cases of joint causation where the contribution of each agent is necessary, but individually not sufficient, to cause the outcome (Spellman & Kincannon, 2001).

1.1.2 Transference theories and omission cases

Transference theories view causation as a physical connection, a process, or a transmission of energy between the cause and the effect. Different accounts have further

articulated this notion to flesh out what the physical connection or the conservation of some physical quantity imply (Dowe, 2000; Salmon, 1984). Here, this very sketchy idea of transference, as contrasted with dependence, will suffice. Transference promises to do better with pre-emption: Peter's behavior is causally relevant (compared to Ed's), and therefore Peter is responsible, because he is the one who is physically assaulting Ted. Transference has proven useful to model psychological causal constructs as well, casting light on how people understand causal relations (Walsh & Sloman, 2005). However, causation cannot always be easily described in terms of physical connectedness. A typical case is that of omissions, where no exchange of physical quantity occurs. Omissions are thus a paradigmatic situation where transference, but not dependence, seems violated (Dowe, 2004).

There is a long-standing debate about whether omissions count as causes, and whether people can be held responsible for allowing harm. Intuitively, there are situations in which allowing harm is morally faulty, and perhaps as faulty as doing harm. As an example, consider these matched scenarios discussed by Rachels (1975): in the *action scenario*, Smith deliberately drowns his cousin in the bathtub. In the *omission scenario*, Jones has a similar plan but fortuitously finds the cousin lying unconsciously in the bathtub and omits to help him. In both scenarios, had the agent behaved differently, the cousin would have survived. No physical transmission took place in the omission one, but dependence explains why we hold both Smith and Jones responsible. Indeed, omitting to help when it would not require too much effort to do so is recognized as reprehensible (Jara-Ettinger et al., 2015). While omissions can perhaps be causes, they are not usually processed in the same way as actions. The omission effect describes the widespread tendency to consider doing harm as morally worse than allowing harm (Anderson, 2003; Cushman et al., 2011; DeScioli et al., 2011;

Zeelenberg et al., 2002). Within our design, we refer to this factor as action initiation (action vs. omission).

Whereas omissions might be treated as causes based on dependence, some theoretical and psychological difficulties persist. These contribute to explaining many instances of the omission effect. To begin with, there is often an asymmetry between actions and omissions in terms of the agent's contribution to the effect (e.g., the existence of Smith, unlike the existence of Jones, is necessary for causing the cousin's death). This is not a necessary feature of omissions though. We can construct omissions where, had the agent not existed, the outcome would not have occurred (Chang, 2009). In philosophy, Kagan discusses a case as such, featuring a dead king and his two sons. After spending his inheritance, the younger son is now starving and unsuccessfully begs his newly-crowned brother, who omits to help, for food. Had the elder not existed, the younger would have become king, thus escaping death (1989). Conversely, there are actions where the existence of one specific agent is not needed for the outcome to occur, but the agent is nonetheless fully responsible for it. In Howard-Snyder's example (2002), the SS officer Franz tortures a prisoner to death. Had Franz never existed, another officer would have carried out the same task.

The under-specification and the causal selection problem are other obstacles to consider omissions as equally causal as actions. The former consists in the difficulty in determining how things would have turned out, had the agent intervened (Wolff et al., 2010): would have Jones successfully rescued his cousin? The latter relates to difficulties in identifying the relevant cause among a set of background omissions (Hesslow, 1988; Livengood & Macherie, 2007). Why is Jones' missing intervention, compared to – for instance – the Queen's, particularly relevant? A solution to this problem can be found at the normative level: omissions are relevant when the agent is

required to intervene. We would not expect the Queen to intervene, but Jones should feel the normative pressure to rescue his cousin (Henne et al., 2017; McGrath, 2005. See also Gerstenberg & Stephan, 2021). One last problem concerns the higher mental cost associated to simulating counterfactuals to omissions compared to actions. Simulating that Jones kills the cousin requires filling in all the details of the murder. Simulating that Smith does not kill his cousin consists in thinking of the same story amended from Smith's action. Byrne and colleagues argue that people more easily consider alternative facts when reflecting upon actions compared to omissions, which results in more causal weight attributed to actions (Byrne, 2007; Byrne & McEleney, 2000).

If the omission bias were simply due to these theoretical and cognitive difficulties, it would not survive their elimination. Stephan and colleagues have shown that clearly individuated omissions are treated as causes, as much as actions, when their mentally simulated alternative has the potential to undo the outcome (Stephan et al., 2017). Analogously, our stimuli were tailored to make it clear when the agent had the potential to successfully undo the outcome, and which counterfactual alternative was relevant. This allowed us to test whether the omission effect was simply driven by contextual difficulties in counterfactual reasoning or rooted in a genuine appreciation of the causal structure of omissions vs. actions.

1.2 Causal pluralism and experimental hypotheses

Dependence accounts easily accommodate the intuition that we are responsible for omissions but have a hard time with pre-emption. By contrast, transference accounts work well with pre-emption but not so much with omissions. If both frameworks are simultaneously taken into account and linearly integrated in responsibility evaluations, we would expect that manipulating them would lead to two effects, i.e., lower

responsibility when transference (omissions) or dependence (pre-emption) fail vs. when transference (action) and dependence (absence of pre-emption) are upheld. This would speak in favor of a pluralistic causal framework applied to responsibility evaluations. Indeed, pluralistic accounts assume that people adopt multiple strategies to conceptualize causal relations. This paper aims to provide further evidence in favor of pluralism in general and, more specifically, to articulate one version of it. One hypothesis (Hp1) is that people alternatively apply dependence or transference depending on the type of scenario. A different hypothesis (Hp2) is that people deploy both dependence and transference intuitions when evaluating the same types of scenarios.

In our main task (Experiment 1), participants were required to evaluate the responsibility of one agent across repeatedly presented, and highly comparable, scenarios. Based on Hp1, the prediction would be that participants apply dependence or transference criteria, but not both, to all presented scenarios. This implies that only one of our manipulations (i.e., dependence or transference) would result in a behavioral effect: if participants apply, consistently and exclusively, transference, then the violation of dependence (i.e., pre-emption) should not affect their responsibility evaluations. Conversely, if participants only apply dependence, then the violation of transference (i.e., omission) should not impact on responsibility ratings. Based on Hp2, transference and dependence are expected to jointly bear on responsibility ratings, thus yielding two main effects, i.e., lower responsibility when dependence or transference or both fail, compared to when they are upheld. The eventual presence of an interaction would speak in favor of an interactive hybrid model where the failure of one type of causal connection has a different impact on responsibility depending on whether the other type of causal connection is upheld or not.

Hp2 is consistent with Chang's hybrid method of making causal attributions (2009), according to which dependence and transference intuitions jointly contribute to the perceived causal role of an agent. In one of Chang's experiments, participants are presented four matched scenarios where an agent (Jim) plays with a model train set, and then asked to evaluate Jim's causal relevance. In each scenario, a house of cards is placed on the track. In scenario1 (dependence and transference upheld), Jim pushes the train, which hits the house. In scenario2 (transference upheld, dependence violated), everything is the same, but a second train approaches the house, hitting it at the same time as the first train does. In scenario3 (dependence upheld, transference violated), Jim lifts a gate to let an upcoming train approach the house. In scenario4 (dependence and transference violated), Jim lifts the gate to let the train pass, while a second train is approaching the house and hits it at the same time as the first does. Whereas the violation of dependence impacts on causal ratings (higher when dependence is upheld), Chang observed no effect of transference and no interaction between the two factors, concluding that dependence overrides transference. The solution contrasts with Chang's hybrid hypothesis and also with Walsh and Sloman's view that transference overrides dependence (2005). One potential concern is that, in scenario3, transference is after all still upheld in virtue of Jim's lifting the gate (action). This differs from Rachels' omission scenario where no transmission occurs, and may help explaining Chang's results.

Building on this, this study compares transference and dependence intuitions as applied to four, repeatedly presented, scenarios: *Action & Preventable Outcome* (A&PO) where dependence and transference are upheld; *Action & Non-Preventable Outcome* (A&NPO) where transference is upheld and dependence is violated; *Omission & Preventable Outcome* (O&PO) where dependence is upheld and transference is

violated; *Omission & Non-Preventable Outcome* (O&NPO) where dependence and transference are violated (Table 1).

[Table 1 near here]

In Experiment 1, we tested the impact of these variables on responsibility ratings. We expected to elicit the omission effect, with lower responsibility in O&PO and O&NPO vs. A&PO and A&NPO, and the pre-emption effect, with lower responsibility in A&NPO and O&NPO vs. A&PO and O&PO, and a possible interaction between the two factors which would show whether the main effects are integrated in a linear fashion or to weighted degrees.

In Experiment 2 and Experiment 3 we tested the impact of the omission and the pre-emption effect on the related dependent variables of intention and control. These notions were selected to track constitutive elements of responsibility, respectively its more mental and its more mechanistic aspect, approximately corresponding to *mens rea* and *actus reus* in the law (Hart & Honoré, 1985; Moore, 2009). This allows investigating how the violation of dependence and transference may differentially affect responsibility with respect to cognate notions that are usually taken into account in evaluations of people's behaviors.

In particular, in Experiment 2, we tested the impact of these variables on intention ratings. Evaluations of the agent's mental states are central to responsibility judgments (Cushman, 2008; Monroe & Malle, 2017; Plaks et al., 2009; Young et al., 2007) both in moral psychology (Mele & Sverdlik, 1996) and in the law (Hart & Honoré, 1985). Peter would not be blamed or punished for unintentionally harming Ed as much as for intentionally doing so. Previous research has emphasized that intention

evaluation is also subjected to the omission effect, in the sense that there is a widespread tendency to consider actions as more intentional than omissions. One explanation for that is that engaging in effortful behaviors tends to be perceived as more intentional than letting an ongoing causal sequence progress towards a predicted outcome (Spranca et al., 1991). We thus expected the omission effect to be present for intentions, with lower intention ratings for O&PO and O&NPO vs. A&PO and A&NPO. By contrast, information about outcome preventability should be irrelevant for evaluating intentions. Whether Ed, unbeknownst to Peter, will hurt Ted in case Peter behaved differently is supposedly irrelevant for evaluating Peter's intention in assaulting Ted. Therefore, we did not expect a significant difference between A&PO and A&NPO and between O&PO and O&NPO. We tested this as the key difference between responsibility and intention evaluations.

Finally, in Experiment 3, we tested participants' evaluations of the very same scenarios by asking them to rate the agents' control over the outcome. This notion of control was selected in order to study the mechanistic and morally neutral aspects of the causal relationships between agents across scenarios. We expected control ratings to match, i.e., go in the same direction as, responsibility ratings as both judgments rely on analogous causal judgments.

1.3 Stimuli

Moral judgments are complex processes supported by different cognitive mechanisms (Sinnott-Armstrong & Wheatley, 2012). Factors like outcome severity (Alicke, 1992; Walster, 1966) or uncertainty about the agent's role and understanding of the situation (Knobe, 2003; Malle, 2001) contribute to modulating responsibility attributions. Responsibility attributions have been traditionally investigated with vignettes describing scenarios that only differ regarding the manipulated factor, e.g., action vs. omission

(Baron & Ritov, 1994). However, such vignettes of everyday situations are saturated with social-contextual information and norms, both known to impact moral evaluation (Willemsen & Reuter, 2016). To exclude possible confounds and to avoid the semantic and social load of vignettes, we employed novel, non-semantic, animated stimuli that depicted the experimental conditions in a purely visual way. Indeed, a number of recent studies attempted to model how visual kinematic events representing interactions may impact causal and even moral judgments, by using non-verbal, visual stimuli or integrating visual and verbal stimuli (De Freitas & Alvarez, 2018; Iliev et al., 2012; Nagel & Waldmann, 2012; Stephan, Mayrhofer, & Waldmann, 2020).

The stimuli were inspired by Michotte's (1946) launching sequences representing causal relations as visual events. Stimuli presenting apparent collisions between objects are widely used in the field of visual cognition to investigate impressions of causality (see, e.g., White's work on generative transmission (2015)). In our design, geometrical shapes (the agents) were connected by colored pipes, used by the shapes to send meaningful colored impulses to a receiving shape (the victim) in the middle of the set-up. Causal interactions were thus represented as visual sequences of delivered impulses, while causal relevance was conveyed by the specific arrangement of shapes and pipes. Working with abstract visual connections allowed us to isolate the causal variables while balancing visual complexity and avoiding reference to intentions or social norms. An additional advantage was that visual stimuli avoided temporal ambiguity regarding the succession of events, which left less room for participants to fill the gaps. In this sense, we aimed at studying the impact of causal events, rather than causal language, on participant's ratings. We made the agent's eventual intervention non-costly, eliminated uncertainty about the agent's understanding of the situation, and clarified the causal contribution of each agent. Furthermore, we counterbalanced the

increased number of agents normally involved in pre-emption by having an equal number of agents in all scenarios.

2. Materials and methods

The orthogonal manipulation of the within-subject factors of action initiation (action vs. omission) and outcome preventability (preventable outcome vs. non-preventable outcome) resulted in four different conditions/types of animations (Figure 1): (a) *Action & Preventable Outcome* (A&PO): the target shape intervenes to bring about the outcome, which would not have occurred without its intervention, i.e., the target shape could have prevented the outcome; (b) *Omission & Preventable Outcome* (O&PO): the target shape does not prevent the outcome, which would not have occurred with its intervention, i.e., the target shape could have prevented the outcome; (c) *Action & Non-Preventable Outcome* (A&NPO): the target shape intervenes to bring about the outcome, which would have occurred even without its intervention, i.e., the target shape could not have prevented the outcome; (d) *Omission & Non-Preventable Outcome* (O&NPO): the target shape does not prevent the outcome, which would have occurred even with its intervention, i.e., the target shape could not have prevented the outcome. Our dependent measures were third-party ratings of responsibility (Experiment 1), intentionality (Experiment 2), and control (Experiment 3).

2.1 Participants

Participants were recruited from Amazon Mechanical Turk and paid $4.50 for participation. Informed consent was obtained, and participants' data were fully anonymized. To reduce variability, participation was restricted to users based in the USA. An a priori sample size of N = 68 was determined in order to detect a

psychologically relevant medium-to-low main effect and interaction in the 2 x 2 design (G*Power, d = 0.4, alpha = .05, power = .90) (Faul et al., 2007).

2.2 Procedure

Participants viewed short animations, representing the four conditions of interest, via their computer web browser (see Supplementary materials for samples of the animations). These animations showed shapes interacting by sending harmful (red), neutral (grey) or inhibitory (blue) impulses via pipes that connected them in various spatial arrangements. The stimuli were designed to depict different causal connections while keeping contextual and semantic information to a minimum. The four causal structures were reproduced multiple times while varying the positions and roles of individual shapes.

Participants were familiarized with the animations in a 5-minute instructions video (see Supplementary materials). They learned that the central grey shape (always a circle) was always the passive receiver, while the surrounding light blue shapes (a star, a triangle, and a square) were active and could deliver impulses. Participants were told that it was up to the shapes to decide whether to deliver the impulse or not through the pipe they were connected to. Depending on the color of its pipe, the shape could act differently: with a red pipe, a shape could send a painful stimulus to the circle, resulting in the circle's receiving a shock. If the stimulus was delivered, participants saw a red light progressing from the shape towards the circle. The circle then turned red, representing pain. In all cases where the circle turned red, all the other impulses immediately stopped. Three different strengths of pain (low/medium/strong, represented by the circle turning light/medium/bright red) were used to introduce visual variation and reduce trial predictability. In contrast, grey pipes could only be used to transmit a neutral grey impulse, which did not affect the circle in any way. If the stimulus was

delivered, participants saw a grey light progressing from the shape towards the circle. Finally, some active shapes were attached to blue pipes connected to another red pipe rather than directly to the circle. By using the blue pipe, the shape could transmit a blue impulse, blocking any approaching red impulse coming from the pipe to which it was attached. If the stimulus was delivered, participants saw a blue light progressing from the shape towards the pipe to which the shape was attached. The instructions clarified that all shapes were ignorant regarding the activity of the other shapes, with the exception of the shapes with blue pipes, which were informed about whether the attached red pipe was active or not and could act fast enough to prevent the shock from reaching the circle. All cases in which the blue impulse was sent were successful, teaching the participant that shapes with blue pipes could always reliably stop the red impulse.

[Figure 1 near here]

Varying in accordance with the dependent measure, the instruction video then introduced the question that participants should answer. For responsibility: "To what extent was the [shape: star/triangle/square] responsible for the circle getting shocked?" (Experiment 1). Participants answered by double-clicking on a visual analogue scale with the extremes labelled "not at all responsible"/"fully responsible". For intention: "To what extent did the [shape] intend the circle getting shocked?" (Experiment 2) on a scale labelled "no intention at all"/"maximal intention". For control: "To what extent did the [shape] control whether the circle got shocked?" (Experiment 3), on a scale labelled "no control at all"/ "full control". Apart from the labels, the three scales looked identical. To avoid a carryover effect, different sets of participants took part in each experiment.

As the question of responsibility for a shock is not defined without a shock, we introduced an opt-out button: when no shock was delivered, participants were instructed to press a "there was no shock" button presented under the response scale. This was further used as an orthogonal measure of attention. We excluded participants for inattentiveness if they failed to press the button more than 10 out of 14 cases when no shock occurred. For the measures of intention and control, participants were instructed to press the "all shapes remained inactive" button when all shapes remained inactive. This happened 4 times and participants who provided less than 3 out of 4 times correct categorizations were excluded as a means of ensuring participant engagement. In four randomly selected trials per condition, after expressing their rating, participants were further asked to explain their judgment by answering an open "Why?" question. These responses were not quantitatively analyzed but served as a check of intelligibility and motivation of the participants.

Participants were instructed that they were allowed to replay each animation as often as they wanted before indicating their rating. After watching the instructions video, participants proceeded to the main experiment: 40 experimental trials (10 per condition: A&PO, O&PO, A&NPO, O&NPO) were presented in random order and interspersed with 50 trials in which the activity of the shape (active/inactive) per pipe-color (red/blue) was counterbalanced in order to minimize prior expectations about the probability of a shape acting or not.

3. Results

3.1 Experiment 1: Responsibility

We collected a total of 83 data sets on MTurk. Predefined exclusion criteria were established: (1) a minimum of 11 out of 14 trials in which no shock occurred were rated

appropriately ("no shock" button pressed) and (2) no more than 10 trials had reaction times above 20 seconds. These criteria led to the exclusion of 15 out of 83 data sets (18%). The remaining 68 data sets (32 female, mean age 37, range 20-63) were used for further analysis.

A 2 × 2 repeated measures with the factors action initiation (action vs. omission) and outcome preventability (preventable outcome vs. non-preventable outcome) yielded a main effect of action initiation ($F(1,67) = 150.72$, $p < .001$, $\eta_p^2 = .69$), a main effect of outcome preventability ($F(1,67) = 17.56$, $p < .001$, $\eta_p^2 = .21$) and an interaction ($F(1,67) = 6.91$, $p = .011$, $\eta_p^2 = .10$). Following up the interaction revealed a stronger effect of outcome preventability in cases of omission ($t(67) = 3.91$, $p < .001$) compared to cases of action ($t(67) = 2.53$, $p = .014$). See Figure 2.

3.2 Experiment 2: Intentionality

We collected a total of 68 valid data sets on MTurk (24 female, mean age 37, range 21-63 years). Predefined exclusion criteria were established: (1) a minimum of 3 out of 4 trials in which no shape acted were rated appropriately ("nothing happened" button pressed) and (2) no more than 10 trials had reaction times above 20 seconds (16 datasets were excluded). Participants had not taken part in Experiment 1.

A 2 × 2 repeated measures with the factors action initiation (action vs. omission) and outcome preventability (preventable outcome vs. non-preventable outcome) yielded a main effect of action initiation ($F(1,67) = 29.99$, $p < .001$, $\eta_p^2 = .31$), while the main effect of outcome preventability ($F(1,67) < 1$, $p < .41$, $\eta_p^2 = .01$) and the interaction ($F(1,67) < 1$, $p = .76$, $\eta_p^2 = .001$) did not reach significance. See Figure 2.

3.3 Experiment 3: Control

We collected a total of 68 valid data sets on MTurk (31 female, mean age 34, range 19-65 years). Participants had not taken part in Experiment 1 or Experiment 2. The exclusion criteria were the same as in Experiment 2 (10 datasets excluded).

A 2 × 2 repeated measures with the factors action initiation (action vs. omission) and outcome preventability (preventable outcome vs. non-preventable outcome) yielded a main effect of action initiation ($F(1,67) = 79.35$, $p < .001$, $\eta_p^2 = .54$), a main effect of outcome preventability ($F(1,67) = 55.47$, $p < .001$, $\eta_p^2 = .45$) and an interaction ($F(1,67) = 28.88$, $p < .001$, $\eta_p^2 = .30$). Following up the interaction indicated a stronger effect of outcome preventability in cases of omission ($t(67) = 7.79$, $p < .001$) compared to cases of direct action ($t(67) = 4.50$, $p < .001$). See Figure 2.

[Figure 2 near here]

3.4 Individual response patterns

While overall rating patterns showed significant effects of both dependence and transference, we were further interested in participants' individual response patterns. Based on the suggestion of an anonymous reviewer,[1] we conducted an exploratory analysis that classified individuals into the categories "dependence-only", "transference-only", "pluralistic" (showing sensitivity to both types of relations) and "none" (if neither a transference nor a dependence rating pattern could be observed) depending on their individual rating responses. For the purpose of classification, first individual judgment difference scores for action vs. omission and preventable vs. non-preventable

[1] We would like to thank both anonymous reviewers for pushing us to think further about our data and initiate this important additional analysis.

conditions were calculated. If a participant showed a difference score of >5 between actions vs. omissions, they were classified as dependence-sensitive. Similarly, a rating difference score >5 for preventable vs. non-preventable outcomes led to the classification as transference-sensitive. The cut-off of ±5 was implemented to account for the inherent imprecision of visual analogue scales. Accordingly, difference scores lower than 5% of the overall scale length (i.e., difference scores between -5 and 5) were treated as insignificant, thus showing no sensitivity to either of the causal concepts. In a last step, participants who were sensitive to only one category were labelled as "dependence-only" or "transference-only", while participants showing differences on both scales were categorized as "pluralistic". Participants who did not show any relevant differences on either scale were labeled as "none". In Experiment 1, this analysis revealed that 25% of all participants applied a pluralistic causal framework when judging responsibility. 63% percent of participants were classified as transference-only raters, while 1% applied dependence-only, and 10% did not show any systematic differences. For intention ratings in Experiment 2, 12% of all participants showed a pluralistic response pattern, 43% showed a transference-only pattern, 4% showed a dependence-only pattern, and 40% did not show any relevant causal sensitivity. For control ratings in Experiment, 3, 51% of all participants showed a pluralistic rating pattern, 29% were sensitive to transference-only, 16% showed a dependence-only pattern and 3% did not show any systematic causal sensitivity. Figure 3 illustrates the proportion of participants that showed sensitivity to either of the causal categories, split by experiment.

[Figure 3 near here]

4. Discussion

By capitalizing on the omission (action vs. omission) and the pre-emption (preventable outcome vs. non-preventable outcome) effect, this study investigated the impact of dependence and transference intuitions on participants' ratings. The stimuli presented interactions between shapes (the agents) delivering colored impulses to harm or rescue the circle (the victim). Overall, we hypothesized that participants' evaluations would have tracked the underlying variations in the causal scenarios shown by the animations, providing evidence for the joint impact of transference and dependence intuitions.

In Experiment 1, participants made responsibility judgements. We observed both the pre-emption and the omission effect. The highest responsibility was assigned when both transference and dependence were upheld, i.e., in A&PO, where the agent did harm and could have prevented it. The lowest when both transference and dependence were violated, i.e., in O&NPO, where the agent allowed harm without being able to prevent it. Reflecting the pre-emption effect, lower responsibility was assigned in A&NPO and O&NPO (dependence violated) vs., respectively, A&PO and O&PO (dependence upheld). Reflecting the omission effect, lower responsibility was assigned in O&PO and O&NPO (transference violated) vs., respectively, A&PO and A&NPO (transference upheld). Taken together, these results speak in favor of the joint contribution of dependence and transference to the evaluation of each scenario.

However, more fine-grained analysis at the level of single participants reveal that many subjects were only sensitive to violations of transference, while a smaller proportion of them actually displayed a pluralistic outlook and almost nobody only cared about violations of dependence. This result is in line with Walsh and Sloman's finding that the presence of a mechanism of transmission from the cause to the effect is particularly central to the causal attributions that underlie responsibility evaluations (2011).

Other contributing factors might, however, explain the omission effect as a violation of proper dependence without referring to a violation of transference. These include people's difficulty in seeing the alternative to an omission as relevant (Phillips et al., 2015), likely to happen (Petrocelli et al., 2011), or to overturn the outcome (Wells & Gavanski, 1989). To test whether the omission effect is truly grounded in a violation of transference, it is crucial to sidestep these factors. Analogously, to single out the pre-emption effect, it is important to avoid possible confounds typical of multi-player environments, including ambiguities about the epistemic status of individual agents or about which agent caused the effect (Li et al., 2011).

Based on our results, the omission and the pre-emption effect are present even when such ambiguities are resolved. The instructions and the stimuli clarified the actual (what one does) and potential (what one can do) causal role of each agent as well as their individual epistemic status. In particular, none of the agents was aware of the others' activity, except for blue pipe-shapes (informed about whether a shock was delivered through the attached red pipe) that could decide whether to rescue the victim. This is crucial to prevent participants from thinking that a blue pipe-shape could lack information about whether the victim was endangered, but also that a red pipe-shape could have information about whether the victim would have received a shock independently of its behavior. This suggests that the effects do not simply stem from contextual difficulties, but rather depend on appreciating differences in the causal structure of events. Overall, responsibility is thus weakened when it is based only on transference but not dependence. Responsibility is also weakened when it is based only on dependence but not transference. By contrast, responsibility is preserved when it is based on both dependence and transference together.

Since both dependence and transference fail in O&NPO, one may expect that the responsibility attributed to the blue pipe-shape here should be at zero. We are not held responsible for omitting to prevent negative events over which we are unlikely to have any impact. For example, although we might be praised for engaging in pacifist efforts independently of their success, we are not individually held responsible for not preventing wars. By contrast, our results show that responsibility here aligns around the midpoint of the scale. We interpret this as follows: blue pipe-shapes only knew about the behavior of the shape they were attached to, being ignorant about the others' activity. As a result, blue pipe-shapes in O&NPO had the same epistemic status as blue pipe-shapes in O&PO, i.e., both lacking information about whether the outcome would have occurred anyway independently of their behavior. Based on their knowledge, they both had an identical opportunity to rescue the circle. Missing intervention suffices to convey a minimal feeling that the agent is involved in the causal sequence. This suggests that participants' judgments rely also on considerations that were not fully operationalized in our design. To check whether responsibility is further diminished, further experiments could investigate trials in which blue-pipe shapes do not intervene while knowing that they cannot prevent the outcome.

The results of Experiment 1 also show a significant interaction between action initiation and outcome preventability, with a stronger effect of outcome preventability in omissions vs. actions. While we have no conclusive explanation of this interaction, a tentative suggestion is that this is driven by a confirmatory predictive mechanism: being in the mind-set of judging omissions, where transference is violated, participants are increasingly sensitive to the presence of other factors further reducing the agents' causal impact. By contrast, when judging actions, where transference is upheld, participants are less sensitive to possible violations of dependence. In other words, when the causal

connection is solidly established at the level of transference, violations of dependence are perceived as relatively less important. Here our results might even tentatively speak for a hybrid causal account, suggesting that the need to represent both causal connections within one scenario can make them interact and mutually constrain each other (Waldmann & Mayrhofer, 2016). However, responsibility in A&PO and A&NPO is relatively high, which makes it impossible to exclude a ceiling effect driving the interaction. Nevertheless, we observed a similar interaction pattern in Experiment 3, where a ceiling effect is less plausible.

A potential caveat before concluding that the results support an overall pluralistic causal model (albeit with important differences at the inter-individual level revealing the predominant role of transference) is the following. One may argue that the stimuli were more apt to elicit a given type of causal representation, i.e., dependence or transference. Indeed, we propose multiple instances of teleological behavior, with agents pursuing goals by doing or allowing harm. Although we did not specifically instruct participants about the sense in which they had to interpret the notion of responsibility, purposefully inflicting pain is commonly thought to be relevant from a moral point of view. Based on this, we interpreted our measurement as tapping into moral, and not just causal, responsibility in teleological scenarios.

In supporting causal pluralism, Lombrozo (2010) plausibly contends that people preferentially, although non exclusively, apply dependence or transference criteria depending on whether a given scenario is teleological or mechanistic. In particular, teleological scenarios more easily elicit dependence intuitions. When people engage in goal-directed behavior, the specific mechanism through which the outcome is achieved is classified as relatively unimportant. We might not care too much about whether Peter harmed Ted with a knife or a gun, as long as the effect is similar. By contrast,

mechanistic scenarios are suited to elicit both transference and dependence. Applying this analysis to our teleological scenarios may suggest that participants were more prone to dependence, compared to transference, intuitions. Further inquiry is needed to determine the proportion of dependence and transference intuitions at play across scenarios. Overall, the relative effect size of the omission and the pre-emption effect as well as the individual response patterns suggest that the failure of transference is more important for responsibility than the failure of dependence. Furthermore, the interaction suggests that both factors are not merely integrated in an additive, linear fashion: when transference is satisfied, the failure of dependence has a lower impact on responsibility than when transference is violated. As such, transference, coinciding here with the ability to initiate a causal sequence, seems to diminish the relative influence of dependence, i.e., the possibility to prevent the outcome, on responsibility.

In Experiment 2, we focused on how differences in action initiation and outcome preventability affect the perception of individual mental states, while keeping contextual information (i.e., outcome valence, agent's epistemic access) homogenous. Different research traditions emphasized the role of intention evaluations in reasoning about responsibility (Cushman, 2013; Lagnado & Channon, 2008; Yuill & Perner, 1988) and causality (Kirfel & Lagnado, 2021). However, few studies have focused on the impact of causal manipulations on intention as a dependent measure (Knobe, 2003). Our results confirmed the hypothesis that intention ratings are affected by how the outcome is achieved (action vs. omission), but relatively immune to outcome preventability. Lacking explicit information about the agents' mental states but knowing that it was up to them to intervene, participants made differential inferences about the extent to which an agent wanted a certain outcome to happen. In line with previous findings, the different motivational valence between actions and omissions plausibly stems from the

fact that the former, marked by physical transmission, are perceived as more effortful than the latter (Greene et al., 2009). The omission effect was thus replicated, with lower ratings in O&PO and O&NPO vs., respectively, A&PO and A&NPO. By contrast, the pre-emption effect was absent with no significant differences between A&PO and O&PO vs., respectively, A&NPO and O&NPO. Correspondingly, individual response patterns revealed that more participants were primarily sensitive to violations of transference rather than to violations of dependence. However, a quite high proportion of them discarded both types of *external* causal information as irrelevant to assess individual *internal* mental states.

Experiment 3 was meant to tackle the more mechanistic aspect of the causal relationships between shapes. As a result, we expected a similar result pattern across Experiment 1 and Experiment 3 as both responsibility and control are grounded in the evaluation of the underlying causal relationships, with control potentially being a subset of responsibility. As predicted, participants attributed less control in O&PO and O&NPO vs., respectively, in A&PO and A&NPO. Thus, the strength of perceived control is not just a function of whether the agent was successful in doing or allowing harm, but also of its causal potential. Action and omission are in fact alike in terms of the agent's ability to prevent the outcome: in A&PO, the agent can successfully do harm; in O&PO, it can successfully prevent harm. However, in A&PO and A&NPO the agent has the further opportunity to initiate a new causal sequence by delivering a shock. By contrast, in O&PO and O&NPO the agent has only a veto power, being able to interrupt a pre-existing causal sequence but not to initiate a new one.

The pre-emption effect was also significant, with participants attributing less control in A&NPO and O&NPO (dependence violated) vs., respectively, A&PO and O&PO (dependence upheld). In A&NPO, the agent exerts control over the current

outcome, but (unlike the agent in A&PO) cannot prevent it. Control reaches a minimum in O&NPO (transference and dependence violated), where the agent has only a veto power and cannot prevent the outcome. Overall, the similar patterns of main effects and interaction support the theoretical and empirical claims that both responsibility and control supervene on causal evaluations, at least when people can grasp the underlying causal scaffolding. Individual response patterns revealed that, compared to responsibility and intention attributions, people rely more heavily on causal information when judging control, which we interpreted as the more mechanistic and morally neutral among the three dependent variables we examined. When judging control, most of participants adopted a pluralistic outlook, with both transference and dependence causal intuitions determining the judgments.

To avoid a carryover effect, three different groups of participants took part in the three experiments. More direct comparisons across the three variables would be made possible by within-subject designs. To minimize the impact of habitual social constraints on moral judgments, we used non-verbal stimuli inspired by Michotte's launching sequence (1946) and Heider and Simmel's animated geometric shapes (1944). In this tradition, causal relations are represented as sequences of events suggesting causal connectedness. Indeed, the human tendency to attribute human-like mental states to non-human, stylized, agents is commonly known (Castelli et al., 2002; Tremoulet & Feldman, 2000). A general concern is related to the ecological validity of our paradigm with respect to language-based vignettes. Nonetheless, the coherent pattern of responses across scenarios and its resemblance with classic tests of action initiation and outcome preventability (Baron & Ritov, 1994; Wells & Gavanski, 1989) suggest that participants interpreted the causal variables we introduced as plausibly mimicking similar real-world interactions between agents. We are not, however, in the

position to exclude that the omission effect is due to the interiorization of (possibly implicit) moral norms. For example, according to deontological ethics doing harm is morally worse than allowing harm and should be judged more harshly (Foot, 1967; Kamm, 1994; Quinn, 1989). Our results might suggest that this moral norm is rooted in an inherent difference in the understanding of causal processes, rather than in contextual difficulties in reconstructing the relevant causal relations.

5. Conclusions

Causal understanding is a permeating aspect of human cognition that allows people to build and update schemas of the environment, including models of others' behavior (Sloman et al., 2009). While causal reasoning has no intrinsic moral flavor, the grasp of causal relations is a key driver in moral assessment. Using a novel visual method, we thus investigated the impact of finely targeted alterations along the cause-outcome dimension of events on responsibility, intention, and control ratings. In the causal scenarios of interest, fully understanding the causal structure required taking into account multiple interacting causal variables. In particular, we observed that responsibility judgments were lowered in cases of omission and pre-emption, with a significant interaction between the two main factors. Based on our results, we proposed that responsibility ratings rely on a combination of dependence and transference intuitions, as applied to analogous types of causal sequences (and with major differences at the interindividual level). Further studies are required to target more specific aspects of responsibility attributions, such as blameworthiness or liability to punishment, or even responsibility for positive outcomes.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*(3), 368–378. https://doi.org/10.1037/0022-3514.63.3.368

Anderson, C. J. (2003). The psychology of doing nothing: forms of decision avoidance result from reason and emotion. *Psychological Bulletin, 129*(1), 139–167. 10.1037//0033-2909.129.1.139

Baron, J., & Ritov, I. (1994). Reference points and omission bias. *Organizational Behavior and Human Decision Processes, 59*(3), 475–498. doi.org/10.1006/obhd.1994.1070

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive and Affective Neuroscience, 11*(2), 138–145. 10.1093/scan/nsw160

Byrne, R. M. (2007). Precis of the rational imagination: how people create alternatives to reality. *Behavioral and Brain Sciences, 30*(5–6), 439–453. 10.1017/S0140525X07002579

Byrne, R. M., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1318–1331. dx.doi.org/10.1016/S0079-7421(08)60501-0

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(8), 1839–1849. https://doi.org/10.1093/brain/awf189

Chang, W. (2009). Connecting counterfactual and physical causation. *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, Austin: TX: Cognitive Science Society, 1983–1987.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: a structural-model approach. *Journal of Artificial Intelligence Research, 22*, 93–115. https://doi.org/10.1007/978-3-642-23963-2_1

Collins, J. (2000). Preemptive preemption. *Journal of Philosophy, 97*, 223–234. 10.2307/2678391

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. 10.1016/j.cognition.2008.03.006

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292. 10.1177/1088868313495594

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive fron nonmoral psychological representations. *Cognitive Science, 35*(6), 1052–1075. 10.1111/j.1551-6709.2010.01167.x

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2011). Judgment before principle: engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience, 7*(8), 888–895. 10.1093/scan/nsr072

De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition, 178*, 133–146. 10.1016/j.cognition.2018.05.017

DeScioli, P., Bruening, R., & Kurzban, R. (2011). The omission effect in moral cognition: toward a functional explanation. *Evolution and Human Behavior, 32*(3), 204–215. doi.org/10.1016/j.evolhumbehav.2011.01.003

Dowe, P. (2000). *Physical causation*. Cambridge University Press.

Dowe, P. (2004). Causes are physically connected to their effects: why preventers and omissions are not causes. In C. Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science* (pp. 189–196). Blackwell.

Faul, F., Erdfelder, E., Lang, A. -G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. 10.3758/bf03193146

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15. 10.1093/0199252866.003.0002

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2012). Noisy newtons: unifying process and dependency accounts of causal attribution. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 523–528.

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 523–528. 10.13140/2.1.3144.2887

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: causal judgments as counterfactual contrasts. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 782–787.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review. Advance online publication.* https://doi.org/10.1037/rev0000281

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: the allocation of responsibility amongst multiple agents. *Cognition*, 166–171. https://doi.org/10.1016/j.cognition.2009.12.011

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition, 216*, 104842. https://doi.org/10.1016/j.cognition.2021.104842

Gilbert, E. A., Tenney, E. R., Holland, C. R., & Spellman, B. A. (2015). Counterfactuals, control, and causation: why knowledgeable people get blamed more. *Personality and Social Psychology Bulletin, 41*(5), 643–658. 10.1177/0146167215572137

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science, 25*(4), 565–610. doi.org/10.1207/s15516709cog2504_3

Greene, G. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371. 10.1016/j.cognition.2009.02.001

Hall, E. (2004). Two Concepts of Causation. In J. Collins, E. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 225–276). Cambridge (Mass): MIT Press.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: a structural-model approach. Part I: causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887. doi.org/10.1093/bjps/axi148

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

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*(2), 243–259. https://doi.org/10.2307/1416950

Henne, P., Niemi, L., Pinillos, Á., & De Brigard, F. (2019). A counterfactual explanation for the action effect. *Cognition, 190*, 157–164. 10.1016/j.cognition.2019.05.006

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: not watering plants. *Australasian Journal of Philosophy, 95*(2), 270–283. doi.org/10.1080/00048402.2016.1182567

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality* (pp. 11–32). New York University Press.

Hitchcock, C. R. (1993). A generalized probabilistic theory of causal relevance. *Synthese, 97*, 335–364. doi.org/10.1007/BF01064073

Hitchcock, C. R. (2003). Of humean bondage. *British Journal for the Philosophy of Science, 54*, 1–25. doi.org/10.1093/bjps/54.1.1

Howard-Snyder, F. (2002). Doing vs. allowing harm. In E. D. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2002 Edition)*. https://plato.stanford.edu/archives/sum2002/entries/doing-allowing/

Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: the role of physical factors in moral judgments. *Memory & Cognition, 40*(8), 1387–1401. 10.3758/s13421-012-0217-1

Jara-Ettinger, J., Tenenbaum, J. B., & Schultz, L. E. (2015). Not so innocent: toddlers' inferences about costs and culpability. *Psychological Science, 26*(5), 633–640. 10.1177/0956797615572806

Kagan, S. (1989). *The limits of morality.* Oxford University Press.

Kahneman, D., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge University Press.

Kamm, F. M. (1994). Action, omission, and the stringency of duties. *University of Pennsylvania Law Review, 142*(5), 1439–1512. 10.2307/3312460

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition, 212*, 104721. https://doi.org/10.1016/j.cognition.2021.104721

Knobe, J. (2003). Intentional action in folk psychology: an experimental investigation. *Philosophical Psychology, 16*(2), 309–324. doi.org/10.1080/09515080307771

Kulakova, E., Khalighinejad, N., & Haggard, P. (2017). I could have done otherwise: availability of counterfactual comparisons informs the sense of agency. *Consciousness and Cognition, 49*, 237–244. 10.1016/j.concog.2017.01.013

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

Lagnado, D. A., Gerstenberg, T., & Zultan, R. I. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 37*(6), 1036–1073. doi.org/10.1111/cogs.12054

Lewis, D. K. (1973a). *Counterfactuals.* Blackwell.

Lewis, D. K. (1973b). Causation. *Journal of Philosophy, 70*. 10.2307/2025310

Li, P., Han, C., Lei, Y., Holroyd, C. B., & Li, H. (2011). Responsibility modulates neural mechanisms of outcome processing: an erp study. *Psychophysiology, 48*(8), 1129–1133. 10.1111/j.1469-8986.2011.01182.x

Livengood, J., & Macherie, E. (2007). The folk probably don't think what you think they think: experiments on causation by absence. *Midwest Studies in Philosophy, 31*(1), 107–127. doi.org/10.1111/j.1475-4975.2007.00150.x

Lombrozo, T. (2010). Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. cognitive psychology. *Cognitive Psychology, 61*, 303–332. 10.1016/j.cogpsych.2010.05.002

Malle, B. F. (2001). Folk explanations of intentional action. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of social Cognition* (pp. 265–286). MIT Press.

McCormack, T., Frosch, C., & Burns, P. (Eds.). (2011). *Understanding counterfactuals, understanding causation: issues in philosophy and psychology.* Oxford University Press.

McGrath, S. (2005). Causation by omission: a dilemma. *Philosophical Studies, 123*(1/2), 125–148. 10.1007/sI 1098-004-5216-z

Mele, A., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies, 82*(3), 365–287. doi.org/10.1007/BF00355310

Menzies, P. (1989). Probabilistic causation and causal processes: a critique of Lewis. *Philosophy of Science, 56*(4), 642–663. doi.org/10.1086/289518

Michotte, A. (1946). *The perception of causality.* Methuen.

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General, 146*(1), 123–133. 10.1037/xge0000234

Moore, M. S. (2009). *Causation and responsibility. An essay in law, morals and metaphysics.* Oxford University Press.

Nagel, J., & Waldmann, M. R. (2012). Force dynamics as a basis for moral intuitions. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 785–790.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology, 100*(1), 30–46. 10.1037/a0021523

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: the relevance of alternative possibilities. *Cognition, 145*, 30–42. 10.1016/j.cognition.2015.08.001

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39*(6), 653–660. doi.org/10.1016/S0022-1031(03)00041-6

Plaks, J. E., McNichols, N. K., & Fortune, J. L. (2009). Thoughts versus deeds: distal and proximal intent in lay judgments of moral responsibility. *Personality and Social Psychology Bulletin, 35*(12), 1687–1701.

Quinn, W. S. (1989). Actions, intentions, and consequences: the doctrine of doing and allowing. *The Philosophical Review, 98*(3), 287–312. doi.org/10.1177/0146167209345529

Rachels, J. (1975). Active and passive euthanasia. *The New England Journal of Medicine, 292*, 78–86. 10.1056/nejm197501092920206

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press.

Scanlon, T. M. (2008). *Moral dimensions: permissibility, meaning, blame.* Harvard University Press.

Schaffer, J. (2003). Overdetermining causes. *Philosophical Studies, 114*(1–2), 23–45. doi.org/10.1023/A:1024457117218

Shaver, K. G. (1985). *The attribution of blame: causality, responsibility, and blameworthiness.* Springer-Verlag.

Shoemaker, D. (2011). Attributability, answerability, and accountability: toward a wider theory of moral responsibility. *Ethics, 121*(3), 602–632. doi.org/10.1086/659003

Sinnott-Armstrong, W., & Wheatley, T. (2012). The disunity of morality and why it matters to philosophy. *The Monist, 95*(3), 355–377. 10.5840/monist201295319

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. doi.org/10.1016/S0079-7421(08)00401-5

Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: experimental findings and implications for jurors' decisions. *Law and Contemporary Problems, 64*, 241–264. 10.2307/1192297

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105. 10.1016/j.jesp.2015.11.005

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation – A computational model. *Cognitive Science, 44*(7), e12871. https://doi.org/10.1111/cogs.12871

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: a computational model. *Topics in Cognitive Science, 10*, 242–257. doi.org/10.1111/tops.12309

Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in inaction: counterfactual simulation and causation by omission. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 1132–1137.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 1–25.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29*(8), 943–951. doi.org/10.1068/p3101

Waldmann, M. R. (Ed.). (2017). *The oxford handbook of causal reasoning.* Oxford University Press.

Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 65, pp. 85–127). Academic Press.

Walsh, C. R., & Sloman, S. A. (2005). The meaning of cause and prevent: the role of causal mechanism. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2331–2336.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: the role of causal mechanism. *Mind & Language*, *26*(1), 21–52. https://doi.org/10.1111/j.1468-0017.2010.01409.x

Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology, 3*(1), 73–79. doi.org/10.1037/h0022733

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*(2), 161–169. doi.org/10.1037/0022-3514.56.2.161

White, P. A. (2015). Visual impressions of generative transmission. *Visual Cognition, 23*(9/10), 1168–1204. doi.org/10.1080/13506285.2016

Willemsen, P., & Reuter, K. (2016). Is there really an omission effect? *Philosophical Psychology, 29*(8), 1142–1159. doi.org/10.1080/09515089.2016.1225194

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: how absences cause events. *Journal of Experimental Psychology: General, 139*(2), 191–221. 10.1037/a0018129

Woodward, J. (2008). Mental causation and neural mechanism. In J. Hohwy, & J. Kallestrup (Eds.), *Being Reduced* (pp. 218–262). Oxford University Press.

Yablo, S. (2002). De facto dependence. *Journal of Philosophy, 99*(3), 130–148. doi.org/10.2307/3655640

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of The National Academy of Sciences*, *104*, 8235–8240.

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology, 24*(3), 358–365. 10.1073/pnas .0914826107

Zeelenberg, M., Van Den Bos, K., Van Dijk, E., & Pieters, R. (2002). The inaction effect in the psychology of regret. *Journal of Personality and Social Psychology, 82*(3), 314–327. doi.org/10.1037/0022-3514.82.3.314

Table 1. Experimental conditions and variables

Figure 1. Experimental design

Figure 2. Results

Figure 3. Individual response patterns

Experimental conditions	Causal frameworks	
	Dependence *Could the agent prevent the effect?*	**Transference** *Is there a physical chain between the cause and the effect?*
Action & Preventable Outcome (A&PO)	YES	YES
Action & Non-Preventable Outcome (A&NPO)	NO	YES
Omission & Preventable Outcome (O&PO)	YES	NO
Omission & Non-Preventable Outcome (O&NPO)	NO	NO

Table 1. Experimental conditions and variables.

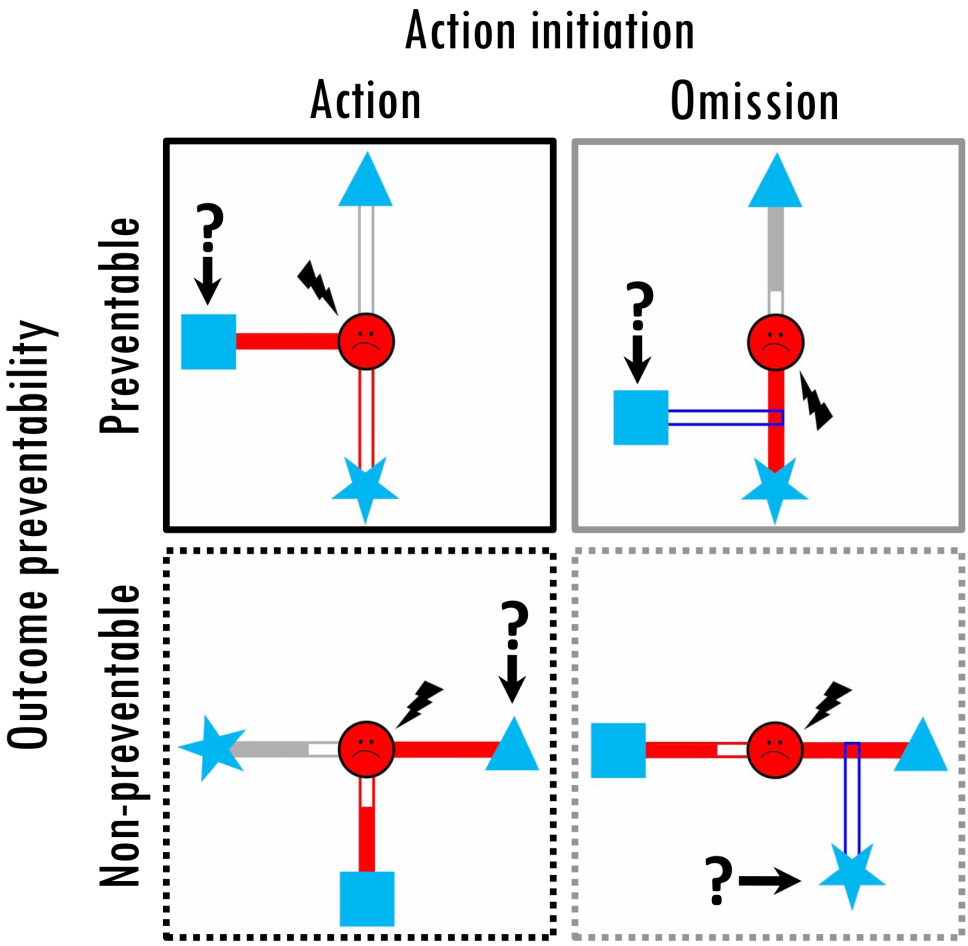

Figure 1. Experimental design. Last frame of animations depicting the four experimental conditions. In each condition, the circle receives a painful shock impulse. The question mark and the arrow, not present in the actual stimuli, here indicate the target shape whose responsibility (Experiment 1), intention (Experiment 2), or control (Experiment 3) participants had to judge. The conditions are arranged as a factorial combination of preventability and action/omission. *Action & Preventable Outcome* (A&PO) (upper left): the square (i.e., the target shape) sends a painful impulse to the circle, making the circle turn red. Had the square not sent this impulse, the circle would not have got shocked; *Omission & Preventable Outcome* (O&PO) (upper right): the

square did not stop the painful impulse sent by the star to the circle. The circle turns red. Had the square stopped this impulse, the circle would not have got shocked; *Action & Non-Preventable Outcome* (A&NPO) (lower left): the triangle sends a painful impulse to the circle that turns red. Had the triangle not sent this impulse, the impulse sent by the square would have shocked the circle; *Omission & Non-Preventable Outcome* (O&NPO) (lower right): the star did not stop the painful impulse sent by the triangle to the circle. The circle turns red. Had the star stopped this impulse, the impulse sent by the square would have shocked the circle. Grey impulses were neutral fillers having no effect on the circle getting shocked.

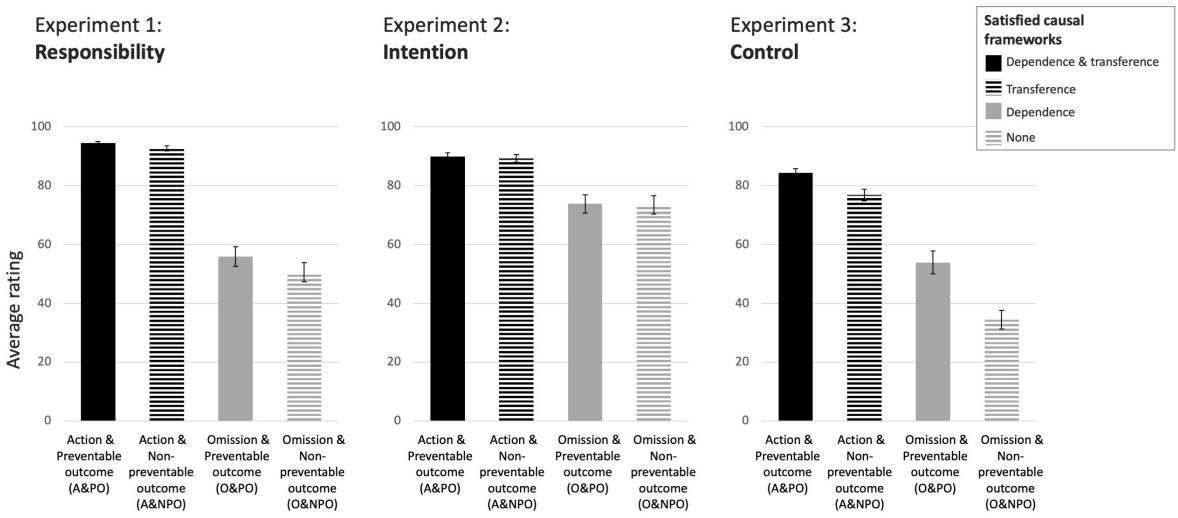

Figure 2. Results. Behavioral results showing the influence of the manipulated factors of action initiation (action/omission) and outcome preventability (preventable outcome/non-preventable outcome) on the dependent measures of Responsibility (Experiment 1), Intention (Experiment 2) and Control (Experiment 3). Error bars indicate standard errors across participants.

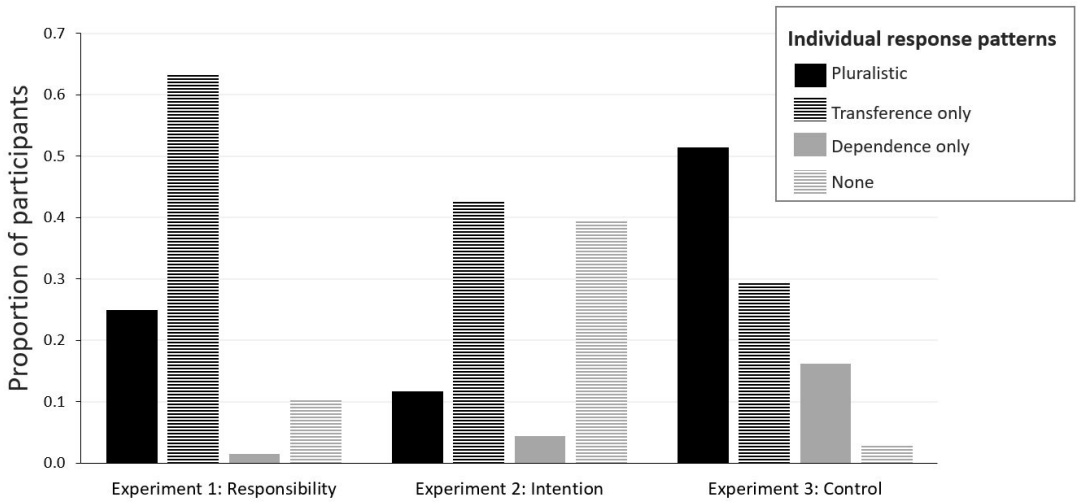

Figure 3. Individual response patterns. Proportion of participants showing only sensitivity to the transference manipulation (Transference only), or the dependence manipulation (Dependence only), both (Pluralistic) or neither (None) in every experiment.