# Attributing Responsibility:
## Actual and Counterfactual Worlds

**Tobias Gerstenberg*** **& David A. Lagnado**
Cognitive, Perceptual and Brain Sciences Department, University College London

February 12, 2013

**Abstract**

How do people attribute responsibility to individuals in a group? Several models in psychology predict that there is a close relationship between counterfactuals and responsibility: how responsible an individual's contribution is seen depends on whether it made a difference to the outcome. We first review these models and then point out a major limitation: people sometimes hold individuals responsible even though their contribution made no relevant difference to the outcome. A richer conception of the relationship between counterfactuals and responsibility is necessary. People's attributions of responsibility are not only influenced by whether a person's contribution made a difference in the actual situation, but also by whether it would have made a difference in other possible situations. We propose a general framework that conceptualizes attributions of responsibility in terms of counterfactuals defined over structured causal models. Using this framework, we show that retrospective responsibility attributions are also affected by prospective responsibility. A person's responsibility depends on how critical their contribution was perceived for the group's success and on how close it was to making a difference to the outcome.

**Keywords:** responsibility; attribution; causality; counterfactuals.

## 1 Introduction

Chelsea played Bayern Munich in the final of the 2012 Champions League – Europe's most prestigious football competition. Bayern scored a goal in the 83$^{\text{rd}}$ minute, but Chelsea's star player, Drogba, equalized in the 88$^{\text{th}}$ minute to carry the teams to extra time. In extra time Bayern's main penalty-taker, Robben, missed a potentially game-winning penalty. Thus, the match went to a penalty shootout. The score in the shootout was at 4–4. Robben had refused the responsibility to take a penalty in the shootout, so it was left to his teammate Schweinsteiger. Schweinsteiger's shot hit the post and he buried his face in his jersey. Now Didier Drogba had the chance to secure Chelsea's first ever victory in the Champions League. He shot, scored and raised his arms to the sky. Chelsea had won. While Chelsea's players celebrated their victory, Schweinsteiger and his teammates were left in tears.

*Corresponding author: t.gerstenberg@ucl.ac.uk, Cognitive, Perceptual and Brain Sciences Department, University College London, WC1H 0AP London.

How do people attribute responsibility to individuals in situations in which several people have collectively contributed to an outcome? We might ask, for example, how much responsibility Robben or Schweinsteiger carry for Bayern's loss in the final. Thoughts about what could have happened readily come to mind: if only Robben had scored his penalty in extra time, Bayern surely would have won the game. If only Schweinsteiger's penalty shot had been an inch to the left he would have scored and Drogba, in turn, might not have been able to cope with the pressure.

In this chapter, we outline a theoretical framework that links attributions of responsibility to counterfactuals. Counterfactual thoughts are thoughts about possible worlds in which the course of events would have unfolded differently from how it actually did (cf. Gerstenberg, Bechlivanidis, & Lagnado, submitted; Kahneman & Miller, 1986; Roese, 1997). Most research in psychology has focussed on retrospective responsibility attributions, where one is concerned with how much responsibility an individual deserves for an outcome that has already occurred (see, e.g., Gerstenberg & Lagnado, 2010; Lagnado & Channon, 2008; Robbennolt, 2000; Zultan, Gerstenberg, & Lagnado, 2012). We might wonder, for example, how responsible Bastian Schweinsteiger is for his team's loss? However, there is another notion of responsibility that is *prospective* and captures the degree to which an individual is responsible for a future outcome. For example, the prospective responsibilities of a defender and a striker are different. It is a striker's responsibility to score goals and a defender's responsibility to prevent them (see Hart, 2008, for a discussion of different notions of responsibility).

In what follows, we will first review previous research that has looked at the relationship between counterfactuals and attributions of responsibility. We will then show how a general framework that conceptualizes attributions of responsibility in terms of counterfactuals defined over causal models illuminates both retrospective and prospective attributions of responsibility (cf. Chockler & Halpern, 2004; Lagnado, Gerstenberg, & Zultan, accepted; McCoy, Ullman, Stuhlmüller, Gerstenberg, & Tenenbaum, 2012). We will conclude by highlighting what we see as the key contributions of our framework.

## 2 Counterfactuals and responsibility

Counterfactual thoughts are commonplace in our everyday life (Kahneman & Miller, 1986; Roese, 1997). If only I hadn't speeded on my way to the office then I wouldn't have got a ticket. If they had invested some more time doing a proper literature review then they surely would have written a better paper. In this section, we will see that there is a close relationship between counterfactuals and attributions of responsibility. How responsible a person is seen for an outcome not only depends on what the person actually did but also on what would have happened if the person had acted differently.

Many counterfactuals can be expressed in terms of counterfactual conditionals with an *if*-part and a *then*-part. *If* Bill had not speeded *then* he wouldn't have got a ticket. Research has shown that attributions of responsibility are sensitive to both parts of the counterfactual (Petrocelli, Percy, Sherman, & Tormala, 2011). Imagine that some nifty neuroscientists found out that people's actions are exclusively determined by unconscious thought processes that are

beyond any explicit control. Presumably, this would have rather substantial consequences for our practices of blaming and praising other people for their deeds (Strawson, 2008; Vohs & Schooler, 2008). The *principle of alternative possibilities of action* is central to many philosophical theories that discuss the relationship between determinism, freedom of will and moral responsibility (Frankfurt, 1969). Should we really blame someone if they could not have acted differently (i.e. if there is no *if*)? While the question of the relationship between freedom of will and moral responsibility is intriguing, we will have to put it aside for the purposes of this essay (see Nichols, 2011; Nichols & Knobe, 2007; Widerker & McKenna, 2006, for more discussion on the topic).

However, even if we assume that people generally have a choice between several courses of action, there is still the question of whether a different action would have actually resulted in a different outcome (i.e. whether the *then*-part of the counterfactual would have proven true). If, for example, it turns out that the same outcome would have prevailed no matter what a person would have done in a particular situation, our first intuition is that the person's responsibility for the outcome is at least diminished if not completely absent. Indeed, "it wouldn't have made a difference anyhow" is a popular excuse when having done something wrong (Kerr, 1996; Kerr & Kaufman-Gilliland, 1997).

In his *culpable control model of blame*, Alicke (2000) highlights the importance of both freedom of choice and the relationship between action and outcome for the blame attribution process. In Alicke's theory, the degree to which a person is blamed for the outcome depends crucially on the perceived control that the person exhibited over their action as well as the outcome. Alicke analyzes control in terms of three structural linkages between mental, behavioral and consequence elements. First, the mind-to-behavior link captures the degree to which the actor's behavior was perceived to have been under volitional control. *Volitional behavior control* is predicted to increase with the belief that the person acted purposefully and knowingly. It is diminished by situational and capacity constraints (we will see an example of each of these constraints below).

Second, the behavior-to-consequence link characterizes to what extent the person had *causal control* over the outcome. Alicke throws in a mix of different criteria that are predicted to influence causal control such as *uniqueness* (How many alternative causes were present?; cf. Kelley's, 1973, discounting principle), *proximity* (How close was the cause to the effect in the chain of events?) and *effective causal control* (What would have happened if the cause had been different?). We will describe different factors that diminish causal control below.

Third, the mind-to-consequences link describes whether the person had *volitional outcome control*. How much volitional outcome control a person has is contingent on whether she foresaw the outcome and desired it. Consider a patient who dies as a result of a wrong treatment by a doctor. In this case, the doctor's volitional outcome control was diminished assuming that the negative consequences of the treatment were in fact not foreseeable and that the doctor did not want the patient to die.

Let us now illustrate in some more detail how causal control (the behavior-to-outcome link) can be weakened. We focus on this notion because it is the most relevant to the general framework of responsibility attribution we will propose in the next section. Consider a slightly

adapted scenario from Wells and Gavanski (1989): Bill goes to dinner with Suzy. It's their first date and Bill, being an old-fashioned gentleman, orders the meal for Suzy. Because the two find themselves in a fancy French restaurant which only serves two different seasonal meals, Bill's choice is somewhat constrained. Bill orders set menu A for Suzy. Unfortunately, the meal contains wine to which Suzy has an allergic reaction and dies.

There are two possible versions of the story: in story one, Bill and Suzy dine at *Restaurant contrôle effective*. In this restaurant, Suzy would have been fine if Bill had ordered meal B. In story two, Bill and Suzy dine at *Restaurant pas de contrôle effective*. In this restaurant, meal B *also* contains wine and Suzy would have also died from eating meal B. So, what's your verdict? In which story is Bill more to blame for Suzy's death? Is he more to blame for having ordered meal A in *Restaurant contrôle effective* or in *Restaurant pas de contrôle effective*? In which restaurant does Bill's choice play a greater causal role? Remember that Bill does the exact same thing in both situations.

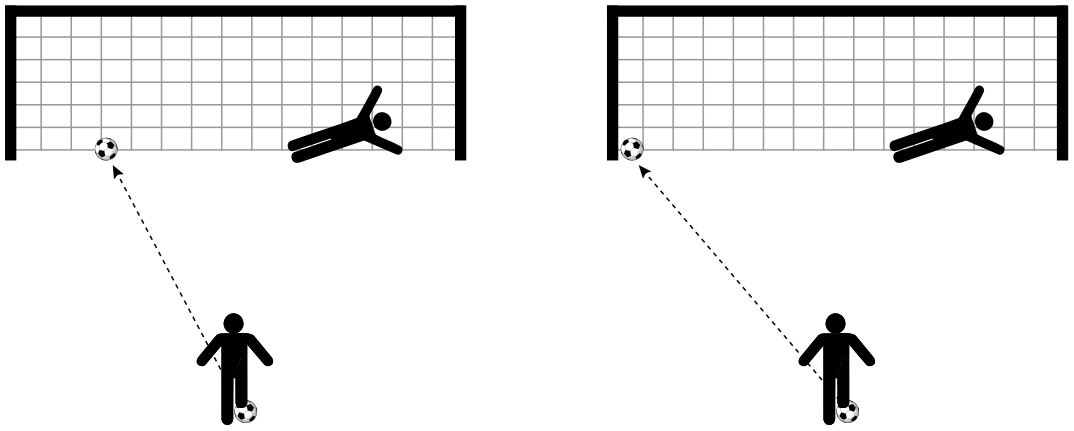
If you think that the causal significance of Bill's choice is greater in *Restaurant contrôle effective*, you are in line with what most participants indicated.[1] If you wonder why Suzy did not tell Bill that she is allergic to wine, you are probably in line with what most participants must have found rather strange about the scenario. In any case, this example illustrates how causal control is diminished if a person finds himself in a situation of no effective control. No matter whether Bill would have chosen meal A or B in *Restaurant pas de contrôle effective*, the outcome would have been the same – Suzy dies. However, in *Restaurant contrôle effective*, Bill had effective causal control (as the name of the restaurant subtly suggests). If only he had chosen meal B, Suzy would have been okay.

As a further illustration, consider the two penalty shots depicted in Figure 1. Let's assume that Bill is now the goalkeeper and Suzy the striker. We know that Bill never saves penalties that are shot very close to the post. However, he gets the penalties that are shot closer to the center of the goal (if he jumps in the correct direction). In both situation (a) and (b), Bill jumped in the wrong direction. In which situation do you think is he more to blame for the goal? The majority of participants in our experiment (67%), who were presented with both situations simultaneously, indicated that Bill is more to blame for not having saved the penalty in situation (a).[2] The rest of the participants indicated that Bill is equally to blame in both situations. Hence, none of the participants thought that Bill was more to blame in situation (b) than in situation (a). In situation (a), Bill could have saved the penalty if only he had jumped in the correct direction. However, the penalty in situation (b) was so well placed that Bill could not have saved it even if he had jumped in the correct direction. Hence, in Alicke's (2000) terms, Bill had effective causal control in situation (a) but not in situation (b).

The restaurant and the penalty scenarios show how both *situational* and *capacity* constraints influence the causal control a person exhibits over the outcome. In one version of the restaurant scenario, the situation was such that no matter what choice the person made, the negative

[1] Wells and Gavanski (1989) did not ask for responsibility judgments in the restaurant scenario but in a conceptual replication in Experiment 2. Causal and responsibility judgments were almost identical.

[2] 21 participants were recruited online via Amazon Mechanical Turk (see Mason & Suri, 2012). 14 indicated that the goalkeeper in (a) is more to blame than the goalkeeper in (b), 7 participants blamed both equally and none of the participants blamed (b) more than (a), $\chi^2(2, N = 21) = 14, p < .001$.



(a) An easy shot to save

(b) A difficult shot to save

Figure 1: How much blame would you assign to the goalkeepers for not having saved the penalty in situation a) vs. b)?

outcome would have resulted. In the penalty scenario, we saw that the goalkeeper's causal control was limited by his capacity. Due to his limited reach, he only has the capacity to save shots that are close to him. In the situation in which the ball is shot close to the post, the goalkeeper could not have saved the ball no matter in which direction he had jumped. If the same outcome prevails no matter what a person does, we feel that their responsibility is diminished – whether Bill dines in the wrong restaurant or happens to have limited goalkeeping abilities.

### 2.1 Theories of retrospective responsibility

In recent years, several psychologists have stressed the importance of counterfactuals for attributions of responsibility. Brewer (1977) was first to propose a model that directly links responsibility attributions with counterfactual considerations. According to her model, people compare the subjective probability of an effect in the presence of the cause $P(E|C)$ with the probability of the effect in the absence of the cause $P(E|\neg C)$. The model predicts that attributions of responsibility are positively related to $P(E|C)$ and negatively related to $P(E|\neg C)$ and that the relationship between each components is additive (see Equation 1). Hence, attributions of responsibility are predicted to be high, only if the probability of the outcome in the absence of the cause is low *and* the probability of the outcome in the presence of the cause is high.[3]

$$Responsibility(C) = P(E|C) - P(E|\neg C) \tag{1}$$

A slightly gruesome but straightforward example is the case of an assassin shooting a victim in the head at point-blank range. Let us imagine that Sarah, Suzy's sister, takes revenge on Bill

[3]Note that the single-event probabilities $P(E|C)$ and $P(E|\neg C)$ should not be confused with the probabilistic relationship between types of events such as in theories of causal learning (Cheng, 1997; Jenkins & Ward, 1965). Rather, $P(E|C)$ is to be interpreted as the subjective degree of belief that the effect will occur in the presence of the cause *in this particular situation*. Similarly, $P(E|\neg C)$ denotes the counterfactual probability that the effect would have come about if the cause had been removed from the actual situation.

for the restaurant episode. Here, the probability of Bill's death given Sarah's shot, $P(E|C)$, is high. Furthermore, under most circumstances we can expect that the probability of Bill's death in the absence of Sarah's shot, $P(E|\neg C)$, is very low. Thus, Sarah is predicted to be judged highly responsible for Bill's death.

Fincham and Jaspars (1983) built on Brewer's (1977) model and added a further component to it. They propose that responsibility attributions are also sensitive to the subjective probability that another person would have acted the same way under the given circumstances (reminiscent of the consensus dimension in Kelley's, 1973, ANOVA model of causal attribution). The more a person believes that others would have done the same, the less responsibility is assigned to the actor. Hence, a revengeful person who believes that most people act in line with the rule "an eye for an eye", would not judge Sarah to be responsible for the outcome. However, a person who thinks that most people would not act in the way that Sarah did, would hold her responsible for what happened.

Spellman (1997) has also taken Brewer's (1977) model as a starting point. In her *crediting causality* model, she extended the Brewer's account in order to handle situations in which an outcome was brought about by a sequence of causal events. For example, Miller and Gunasegaram (1990) described the following scenario:

> Imagine two individuals (Jones and Cooper) who are offered the following very attractive proposition. Each individual is asked to toss a coin. If the two coins come up the same (both heads or both tails), each individual wins $1,000. However, if the two coins do not come up the same, neither individual wins anything. Jones goes first and tosses a head. Cooper goes next and tosses a tail. Thus, the outcome is that neither individual wins anything.

Who do you think is to blame? Jones or Cooper? When presented with a forced-choice format, 92% of the participants indicated that Jones, who tossed first, would blame Cooper more for the negative outcome than vice versa (see also Mandel, 2003).

According to Spellman's (1997) model, participants evaluate the causal contribution of each event in a chain by comparing the probability of the outcome *before* and *after* the causal event occurred. The model predicts that people's judgments of how much an event causally contributed to the outcome is related to how much the event changed the probability of the outcome. The model predictions can be nicely illustrated via the coin-toss scenario (see Equation 2). The probability of Jones and Cooper together winning the $1,000 before any of them tossed their coins is $P(win) = 50\%$. After Jones has tossed head, the probability of them winning is still $P(win|Jones\ tossed\ head) = 50\%$. Jones's toss did not change the probability of the outcome and he is thus predicted not to have causally contributed to the outcome. For Cooper, in contrast, the situation looks different. Before he tossed his coin, the probability of them winning was 50%. However, after he tossed his coin, the probability decreased to 0% (it would have increased to 100% had his coin matched Jones's). Hence, while Jones's action did not change the probability of the outcome, Cooper's did. In a series of experiments in which Spellman (1997) varied the extent to which the probability of the effect was changed by the different causal events in the chain, the crediting causality model predicted people's causal

judgments as well as their attributions of blame very accurately.

$$
\begin{aligned}
Responsibility(Jones) &= P(win|Jones\ tossed\ head) - P(win) \\
&= 0.5 - 0.5 = \mathbf{0} \\
Responsibility(Cooper) &= P(win|Cooper\ tossed\ head\ \&\ Jones\ tossed\ head) \\
&\quad - P(win|Jones\ tossed\ head) \\
&= 1 - 0.5 = \mathbf{0.5}
\end{aligned} \tag{2}
$$

The simplicity of Spellman's (1997) account as well as its broad applicability have made it a very popular model in the attribution domain. However, in recent years, several studies have pointed out limitations of the account. Spellman's models predicts that an event's perceived causal contribution merely depends on how much it changed the probability of the outcome. However, the same change in the probability can come about in many different ways, some of which might strike us as more causal/responsible than others. For example, consider that Bill (who miraculously recovered from the blow of Sarah's shot in the meantime) is on his way back home in his car. As he reaches a curve, the car loses track on a wet patch, comes off the road and crashes into a tree. Bill is severely injured. According to Spellman's model the fact that the road was wet contributed causally to Bill's crash assuming that the probability of the crash would have been lower had the road been dry. The reason for *why* the road was wet, however, is not predicted to influence the perceived causal contribution. For example, a sudden shower of rain is predicted to have no less responsibility for the outcome than Sarah who intentionally sprinkled the street with a hose.

Contrary to this prediction, studies have shown that people attribute more causality and blame to events that have been brought about intentionally rather than accidentally (cf. Lagnado & Channon, 2008; Lombrozo, 2010). People also prefer voluntary human actions over physical causes as explanations for why an outcome occurred (McClure, Hilton, & Sutton, 2007) and are more likely to select voluntary human actions rather than physical events as the causes for a negative outcome, even if the extent to which the probability of the outcome is changed is controlled for (Hilton, McClure, & Sutton, 2010).

Mandel (2003) has shown another important limitation of Spellman's model. People sometimes select an event as the cause of the outcome even if the event did not change the probability of the outcome at all. Consider the following scenario: Bill, who just left the hospital after having recovered from the car accident, is having a beer in his favorite bar. While he is in the bathroom, Sarah who has been following him, pours lethal poison into Bill's beer and immediately sneaks away. After having finished his beer, Bill heads back home thinking to himself that he should try a different beer the next time because of the strange aftertaste in his mouth. On his way he is ambushed by Jack, Suzy's and Sarah's raging brother. Jack, an established member of the local gun club, aims at Bill's head, shoots and hits. Bill falls to the ground – dead.

Who do you think caused Bill's death, Sarah or Jack? Spellman's model predicts that Sarah's poisoning will be selected as the cause for Bill's death. The poison increased the probability of

Bill's death from very low to 100%. However, in a similar scenario employed by Mandel (2003), participants rated the final event which actually brought about the death of the protagonist as causally more important than the first event which had already increased the probability of the effect to almost certainty. One way of rescuing Spellman's model would be by specifying the actual outcome more precisely. If, for example, the question was whether Sarah is responsible for "death by gun shot", the answer would presumably be negative. However, we will see below that the strategy of increasing the granularity of the effect will not always help.

Let us conclude this brief review of theoretical accounts that propose a close relationship between counterfactuals and attributions of responsibility with a recent addition. We have seen above that counterfactuals can be expressed in terms of a conditional *if ... then ...* statement. According to Petrocelli et al. (2011), the extent to which a person is predicted to be held responsible for a particular outcome increases with the *potency* of a relevant counterfactual. A counterfactual's potency is given by multiplying the probability of its antecedent (i.e. the *if-likelihood*) with the probability of its consequent (i.e. the *then-likelihood*, see Equation 3).

$$\textit{Counterfactual Potency} = \textit{if-likelihood} \times \textit{then-likelihood} \tag{3}$$

Petrocelli et al.'s (2011) model is nicely illustrated via the restaurant scenario introduced earlier. The relevant counterfactual is: *If Bill had chosen meal B (instead of meal A) then Suzy would have survived (instead of died).* In order to evaluate how potent the counterfactual is, we need to consider how likely it was that Bill could have chosen meal B *and* that Suzy would have survived given that Bill had chosen meal B. Because the likelihood of the antecedent and consequent are predicted to combine in a multiplicative way, both likelihoods have to be greater than zero in order for the counterfactual to be potent.

In *Restaurant pas de contrôle effective*, the *then-likelihood* is zero. If Bill had chosen meal B, Suzy would still have died. Hence, the potency of the counterfactual is zero and the model predicts that Bill will not be held responsible for the outcome. In *Restaurant contrôle effective*, the *then-likelihood* is high. Suzy would have been ok, had Bill chosen meal B. How responsible Bill is seen in this restaurant now depends on the likelihood of the counterfactual's antecedent. For example, if Bill is described as having been unsure which meal to choose and, after going back and forth a few times, eventually decided to order meal A, then the counterfactual of having taken meal B instead is highly available. However, if Bill is described at having been very sure about going for meal A and did not even consider any other option, then the *if-likelihood* is low and hence his responsibility for the negative outcome is predicted to be low. Petrocelli et al. (2011) demonstrate in a range of studies that their model accurately predicts the qualitative trends in participants' attributions.

## 2.2 A problem case for simple counterfactual theories

All the models discussed so far link counterfactuals and attributions of responsibility in a fairly straightforward manner. Essentially, a person's responsibility is a function of the degree to which a person's contribution actually made a difference to the outcome. Making a difference

is either defined in terms of the degree to which an event changed the probability of the effect (Brewer, 1977; Fincham & Jaspars, 1983; Spellman, 1997) or the potency of a considered counterfactual (Petrocelli et al., 2011). However, in a recent study we have demonstrated that attributions of responsibility are not only influenced by the extent to which a person's contribution made a difference to the outcome in the actual situation but also by whether the person's contribution was likely to have made a difference if the actual situation had turned out to be somewhat different (Gerstenberg & Lagnado, 2012).

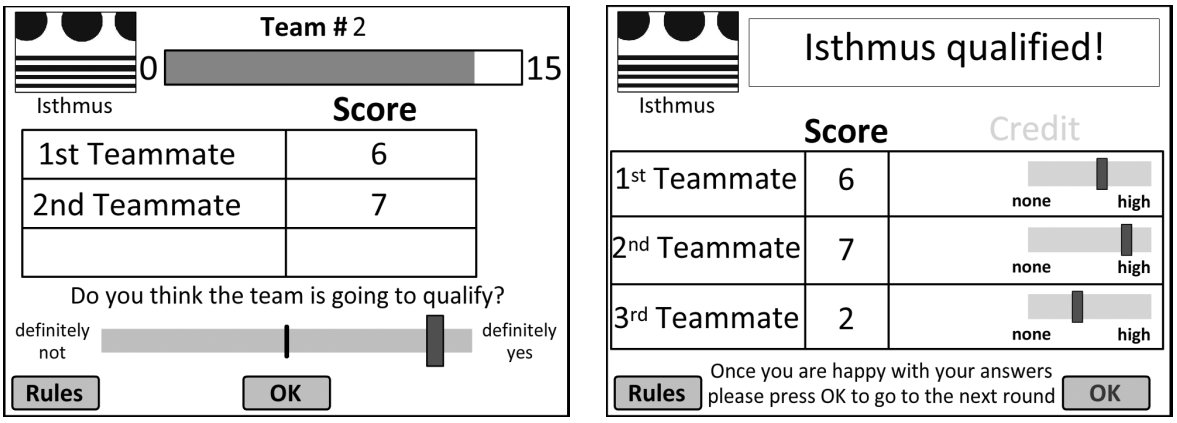
Consider a hypothetical scenario in the men's 4 x 400m team-relay final of the London Olympics 2012. Both Great Britain (GB) and Germany have made it to the final. GB has a very good start and by the time the third German runner passes on the baton to the fourth runner, team GB has already crossed the finishing line. Now consider two versions of how the story could unfold: either the fourth German runner performs very poorly or he performs exceptionally well.

According to Spellman's model, the causal contribution of the final German runner was identical in both situations because the loss of the German team was already determined prior to the final runner (assume, for simplicity, that no other teams competed in the final due to an unprecedented doping scandal). However, the intuition is strong that the German runner will be held less responsible for the loss if he ran well compared to if he ran poorly. Despite the fact that his performance did not matter for the outcome in the actual situation, one can easily think of other possible situations in which the performance of the fourth runner would indeed have mattered. If, for example, the English runners had been slightly slower, a very fast fourth runner could have made the German team win. The slow runner, in contrast, sends an ambiguous signal with his performance (cf. Reeder & Brewer, 1979). It could be that he did not try hard because he already knew that the team's loss was guaranteed. Alternatively, it could be that he is just not a very good runner. If this were true, he would have presumably also made his team lose even if the conditions had been more favorable.

Through considering that other possible situations could have occurred, we can explain why the two runners are treated differently: a good runner would likely have made the team win if the situation had been a bit more favorable whereas a bad runner would have likely caused the team's loss even in a more favorable situation.

In order to test these intuitions about the differences between the two runners, we devised an experiment in which participants evaluated the performance of different teams in a competition (see Gerstenberg & Lagnado, 2012). Each team was comprised of three athletes who performed their individual routines sequentially. Athletes could receive a maximum of ten and a minimum of zero points for their performance. In order for a team to be successful they needed to get a total of 15 points or more.

Participants in our experiment first saw the scores of the three athletes sequentially (see Figure 2a). After each athlete's score, they were asked to indicate how likely they thought the team would reach the qualification criterion of 15 points. Once participants had made their predictions and found out whether or not the team actually qualified, they saw the scores of all three athletes simultaneously and were asked to indicate to what extent each athlete was responsible for the team's outcome (see Figure 2b).




(a) Probability updating phase

(b) Responsibility attribution phase

Figure 2: Screenshots of Gerstenberg & Lagnado's (2012) experiment.

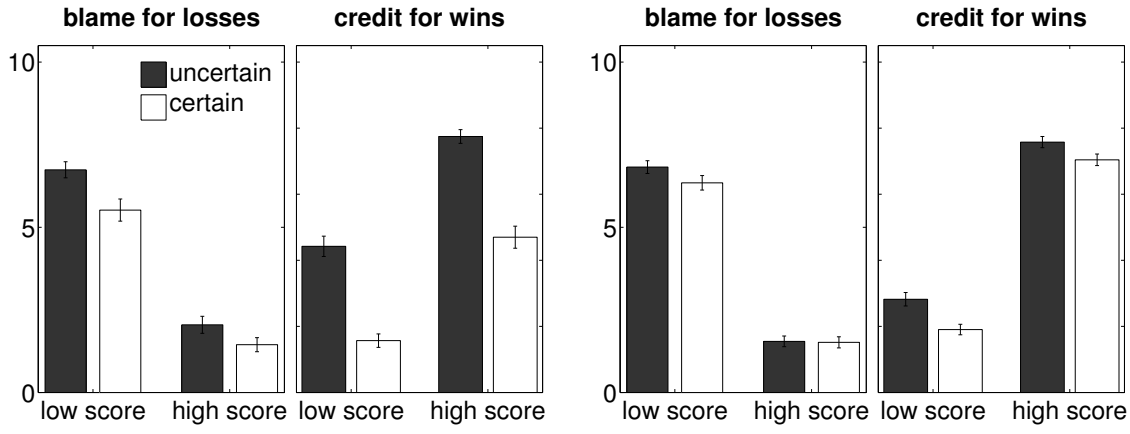
In our analysis, we focused on how much responsibility was attributed to the third athlete in the team as a function of i) how they performed and ii) whether or not the team's outcome was already determined prior to their performance. Our setup allowed for two different ways in which the team's outcome could be certain even before the third athlete competed. If, for example, the performance of the first two athletes was very poor and their individual scores added up to less than 5 points, the team's loss was already certain. There was no way for the third athlete to make the team win as she can only score a maximum of 10 points. Conversely, if the performances of the first two athletes were very good, they could already secure the team's successful qualification prior to the third athletes turn by achieving a combined score of 15 points or more.

The results showed that participants' responsibility attributions to the third athlete were strongly affected by performance (see Figure 3). Generally, the third athlete received more credit for his team's success and less blame for a failure, if his score was high rather than low. Furthermore, in line with Spellman's model, attributions of responsibility to the third athlete were reduced when the team outcome was already certain before he performed compared to when the result was still unsure (see Figure 3a). However, in contrast to the predictions of Spellman's model, participants' responsibility attributions varied with the third athlete's performance *even when* the outcome was already certain.

Importantly, the effect of outcome certainty was only present when participants had been instructed that athletes who performed later knew the scores of their teammates. When participants were instructed that later athletes did not know about the scores of their teammates, the effect of outcome certainty disappeared (see Figure 3b). Hence, whether participants reduced their attributions in the light of the certainty of the team's result dependent on the athlete's knowledge rather than their own knowledge as an external observer.

### 2.3 The need for a richer counterfactual framework

The results of the previous experiment demonstrate that the relationship between counterfactuals and attributions of responsibility is not as simple as some theories have proposed (e.g. Brewer, 1977; Petrocelli et al., 2011; Spellman, 1997). For a person to be held responsible, it




(a) Knowledge condition (b) No knowledge condition

Figure 3: Attributions of responsibility to the third athlete as a function of the athlete's performance (low score / high score) and the certainty of the outcome (uncertain/certain) prior to the athlete's performance. Results are shown separately for negative outcomes (left side of each panel) and positive outcomes (right side of each panel). Error bars indicate ±1 standard error of the mean.

not only matters whether their action made a difference in the actual situation but also whether it would have made a difference had the situation been somewhat different. One way to think about these effects is in terms of the robustness of the counterfactual relationship between the events of interest. Woodward (2006) has argued that people are more willing to say that a particular event $C$ caused another event $E$ when the counterfactual relationship between the two events is insensitive to changes in the background conditions (see also Campbell, 2008). More precisely, a counterfactual is robust to the extent that *even if* the situation had been somewhat different, (i) $E$ would still have happened if $C$ had happened ($C \rightarrow E$) and (ii) $E$ would still *not* have happened if $C$ had not happened ($\neg C \rightarrow \neg E$).

Lombrozo (2010) has provided empirical evidence for the role of robustness in people's causal judgments. In a series of experiments, she showed that people are more willing to judge cases of double prevention (in which one event prevents another event from preventing an outcome) as causal when an agent acted intentionally as opposed to accidentally. If the situation had been somewhat different, an intentional agent would have been able to adjust their behaviour in order to ensure the desired outcome (cf. Heider, 1958). If, in contrast, the outcome was the result of an accident, then a small change in the background conditions would have resulted in a different outcome.

Applied to our findings from the experiment reported above, consider the situation in which the third athlete performs well despite the fact that the team has already qualified. Even though it is true that the team would still have qualified if his performance had been worse in the actual situation, the counterfactual relationship between a good performance and a positive team outcome is more robust than the counterfactual relationship between a *poor* performance and a positive outcome. Assuming a poor performance by the third athlete, minimal changes in the background conditions (such as a slightly worse performance of the other teammates or a minimally higher qualification criterion) would have been sufficient to undo the positive group outcome. However, when the third athlete's score was high, the team would have won even if other relevant factors had been somewhat different.

If causal and responsibility judgments are not only a function of what happened in the actual situation but also of what would have happened in other possible situations, an important question concerns what sorts of other possible situations people are likely to consider. Recent findings in experimental philosophy suggest that people's beliefs about what normally happens influence what counterfactual worlds are deemed relevant (Halpern & Hitchcock, forthcoming; Hitchcock & Knobe, 2009, see also Hart & Honoré, 1959/1985; Kahneman & Miller, 1986). Rather than thinking about what would have happened in some world that is arbitrarily different from the actual world, people tend to consider worlds that are closer in line with moral norms (what should have happened) or statistical norms (what was likely to happen, Hitchcock & Knobe, 2009). Applied to our example, the third athlete should receive particularly high praise for a good performance when the previous athletes performed better than expected. In such a situation, the counterfactual in which the third athlete's good performance would have been necessary to make the team win is highly available. It is easy to imagine that the performances of the previous athletes might have been worse.

But how do people evaluate what would have happened in a given counterfactual world once they have decided which one to consider? In the next section, we will argue that people make use of their causal understanding in order to predict what would have happened under different possible contingencies (cf. Gerstenberg et al., submitted; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012). We will argue that an adequate theory of responsibility attribution needs to be precise about how people represent the causal structure of the situation.

## 3 Causality and responsibility

In the previous section, we have seen how different models predict a close correspondence between the extent to which a contribution made a difference to the outcome and a person's degree of responsibility. Legal theories also employ a counterfactual criterion as a first test to causality. According to the *sine qua non* condition, an action qualifies as a cause if the outcome would not have occurred *but for* the action (Hart, 2008; Hart & Honoré, 1959/1985; Moore, 2009; Spellman & Kincannon, 2001). However, there are also situations in which we would like to hold a person responsible *despite the fact* that their contribution made no significant difference to the outcome. Let us reanimate poor Bill once more for a final execution. Sarah and Jack both shoot Bill simultaneously from point-blank range. Each of Sarah's and Jack's shot would have been sufficient to bring about Bill's death. How responsible is Jack for Bill's death?

According to the *sine qua non* condition, neither Jack nor Sarah qualify as a cause of Bill's death. Even if Jack had not shot, Bill would still have died (because of Sarah's shot). In Alicke's (2000) terminology, Jack lacked effective causal control. The different accounts discussed in the previous section (Brewer, 1977; Fincham & Jaspars, 1983; Petrocelli et al., 2011; Spellman, 1997) would also have to conclude that neither Jack nor Sarah are responsible for Bill's death. Neither of their individual actions made a difference to the probability of the outcome (assuming that it is almost impossible that either of them could have missed their shot). Hence, we are left with a conundrum: a dead person and no one is responsible. However, it seems that both

Jack and Sarah carry (at least partial) responsibility for the outcome. So we have an outright clash between two strong intuitions: on the one hand we think that a person should only be held responsible to the extent that her contribution made a difference to the outcome. On the other hand, we would sometimes like to attribute responsibility to a person *even if* her action made no difference to the outcome such as in situations in which the outcome was causally overdetermined.

In order to resolve this issue, we will need to say a bit more about the relationship between causality and responsibility (up until now we have not really distinguished these concepts). Most theoretical frameworks of responsibility attribution in psychology conceive of the relationship between causality and responsibility in terms of entailment (Darley & Shultz, 1990; Fincham & Jaspars, 1980; Shaver, 1985). That is, causality is a necessary condition for responsibility which in turn is a necessary condition for blame and punishment.[4] Hence, according to these models, an ideal observer should first assess whether the person of interest was the cause of the negative outcome. Additional criteria such as foresight and intention have to be met to qualify the person as being responsible and blameworthy for the outcome (Lagnado & Channon, 2008). However, there have also been accounts that deny this entailment relationship and argue that moral evaluations influence attributions of causality (Alicke, 2000; Knobe, 2010). The fact that neither Jack nor Sarah identify as causes according to a simple counterfactual criterion and that we nevertheless have the intuition that they bear some responsibility for the outcome is problematic for the entailment view of causation and responsibility.

### 3.1 Causal model framework

The theoretical framework of causal networks (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000) has helped to sharpen the debate about causality through providing a common language for philosophers, psychologists and computer scientists. Causal networks consist of a graph and a corresponding set of structural equations which describe the dependencies between the different variables in the network. Figure 4 shows a graph that represents the overdetermination scenario described above.

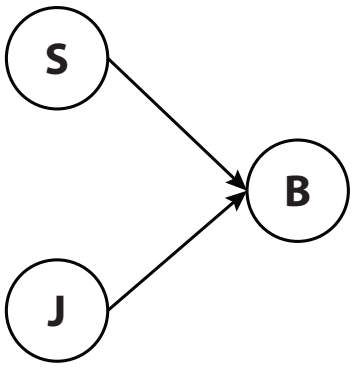



Figure 4: A simple causal network representing the overdetermination scenario. S = Sarah: {0 = doesn't shoot, 1 = shoots}, J = Jack: {0 = doesn't shoot, 1 = shoots}, B = Bill: {0 = survives, 1 = dies}.

[4]The concept of vicarious responsibility is a prominent exception to the strong entailment view such as when a manager is held responsible for the actions of their employees (Gregory, 1932; Shultz, Jaggi, & Schleifer, 1987).

A causal graph consists of a set of nodes which represent the variables of interest and a set of directed edges between the nodes which represent the directionality of causal dependence. For simplicity, we assume that each of the variables in our network is binary. Sarah can either *shoot* ($S = 1$) or *not shoot* ($S = 0$). The same is true for Jack. Bill can either *die* ($B = 1$) or *survive* ($B = 0$). While the directionality of (causal) dependence can be directly read off the graph (e.g. we can see that $B$ is caused by $S$ and not vice versa), the graph structure does not reveal how different causes combine to bring about an effect. The exact relationships between the variables is given by the corresponding structural equations. For our example, we need to specify how $S$ and $J$ determine the value of $B$.[5] Given that each $S$ and $J$ are individually sufficient to bring about $B$, we know that $B = max(S, J)$. That is, $B = 1$ if either $S = 1$ or $J = 1$ or both.

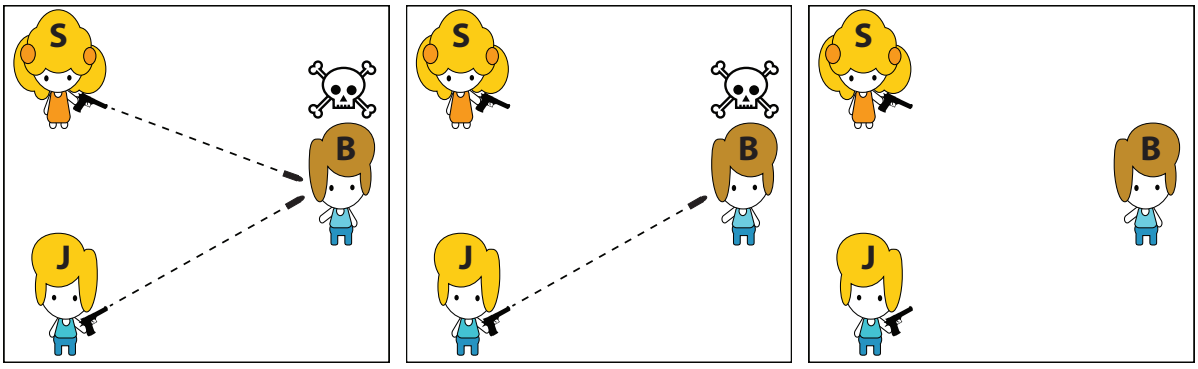
A causal network concisely captures core theoretical concepts that have been argued to be intimately linked with causality, such as prediction, explanation and counterfactuals (Schaffer, 2003). For example, we can predict what happens to $B$ given the value of $S$. We can also reason diagnostically and infer from observing the value of $B$ that either $S$ or $J$ or both have to be positive. Additionally, a causal network also supports reasoning about interventions (What would happen if I forced Jack to shoot?) and about counterfactuals (Would Bill still have died if Sarah hadn't shot?). Finally, we will also see that we can use causal networks to define how much responsibility different variables have for a certain outcome.

The causal model and associated structural equations imply a set of possible worlds (or a distribution over possible worlds when dealing with probabilistic rather than deterministic relationships between variables). Figure 5 shows three of the four possible worlds that are consistent with the causal representation of the shooting scenario. Figure 5a shows what happened in the actual world. We have seen that a simple counterfactual criterion does not identify Sarah or Jack as a cause of Bill's death in this situation. A central idea now is that the notion of counterfactual dependence needs to be relaxed in order for variables to qualify as causes even when they did not make a difference to the outcome in the actual world. Several researchers have offered accounts that define causality in terms of counterfactual dependence under certain contingencies (Halpern & Pearl, 2005; Hitchcock, 2001; Woodward, 2003; Yablo, 2002).

We will focus on the account offered by Halpern and Pearl (2005). According to their account, $J$ qualifies as a cause of $B$ if there is a counterfactual dependence between $J$ and $B$ given that variables that are not on the causal path from $J$ to $B$ are held fixed at a certain value. If we change $S$ from its original value of $1 =$ *shoot* to $0 =$ *not shoot*, we generate a situation in which $B$ now counterfactually depends on $J$. If we fix $S = 0$, Bill would have died if Jack had shot (Figure 5b) but survived if Jack had not shot (Figure 5c).[6] We will use the word *pivotal* to refer to a variable which is in a relationship of counterfactual dependence with the effect. For example, while $J$ is not pivotal in Figure 5a because there is no counterfactual dependence between $J$ and $B$ in that situation, $J$ is pivotal for Bill's death in Figure 5b and

[5] We assume that the values of the root variables $S$ and $J$ are determined by factors that are external to our model representation.

[6] Halpern and Pearl (2005) impose constraints on what variables can be held fixed at certain values in order to check for counterfactual dependence. Because we will deal with fairly straightforward examples, these additional constraints need not worry us here.




(a) Both shoot, Bill dies; Jack's responsibility = 1/2

(b) Only Jack shoots, Bill dies; Jack's responsibility = 1

(c) Neither shoots, Bill survives; Jack's responsibility (for Bill's survival) = 1

Figure 5: Overdetermination scenario with responsibility predictions by Chockler and Halpern's (2004) model. S = Sarah, J = Jack, B = Bill.

for Bill's survival in Figure 5c.

### 3.2 A causal model theory of responsibility attribution

With this machinery in place, we can finally address the problem of how much responsibility individual variables should carry for an effect they brought about collectively. Chockler and Halpern (2004) have provided a model that defines responsibility in terms of the minimal number of changes that have to be made to the original situation, in order to generate a situation in which the variable of interest is pivotal for the outcome. More precisely, the responsibility of a target variable $C$ for an outcome $E$ is given by

$$\textit{Responsibility}(C,E) = \frac{1}{N+1}, \tag{4}$$

where $N$ equals the minimal number of other variables in the causal network whose value have to be changed in order to render $C$ pivotal for $E$. For example, in Figure 5a, Jack is not pivotal. However, if we consider another possible world in which Sarah had not shot, that is, we change the value of one variable ($N = 1$), then Jack is pivotal. Hence, Jack receives a responsibility of $\frac{1}{1+1} = \frac{1}{2}$ for Bill's death in the situation in which Sarah shot as well. Figure 6 shows how the responsibility of a variable changes as a function of how many changes would be necessary to render the variable pivotal.

If the variable is pivotal in the actual situation and hence the number of changes $N = 0$, the variable is fully responsible for the outcome. However, unlike a simple counterfactual model, responsibility does not drop to zero once a variable is not pivotal. Rather, responsibility reduces quickly with the number of changes necessary for pivotality and asymptotes towards zero for large numbers of changes (cf. Latané, 1981, for a similar proposal that the relative decrease in responsibility becomes smaller with an increased number of people).

It is worth noting that due to the very general way in which the model defines responsibility, it applies equally well to situations in which several agents contribute to an outcome or in

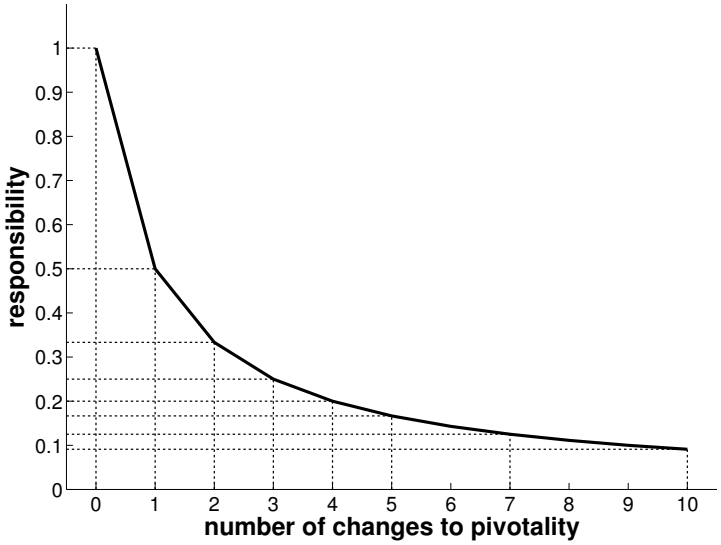



Figure 6: Responsibility function according to Chockler and Halpern (2004).

which several parts of a machine are required to work in order to produce an outcome (see Chockler, Halpern, & Kupferman, 2008). Furthermore, the model assumes that people consider counterfactuals worlds in terms of changes to the values of the variables in the network and *not* in terms of changes of the causal structure of the situation (cf. Chockler & Halpern, 2004). It thus provides a more precise notion of the (causal) similarity between different possible worlds (Pearl, 2000; cf. Lewis, 1973, 2000, for previous attempts).

### 3.3 Empirical tests of the theory

In previous work (Gerstenberg & Lagnado, 2010), we have tested the predictions of Chockler and Halpern's (2004) structural model of responsibility attribution. Participants' task in our experiment was to count the number of triangles in briefly presented complex geometrical forms. Participants were paired up with three virtual players to form a team. Between conditions, we manipulated the way in which the individual performances combine to determine the team's outcome (cf. Steiner, 1972). In the *sum* condition, each individual player's error was added and the team succeeded if their summed individual errors were less than 7. A player's error equalled the deviation from the correct solution. For example, if the diagram contained 23 triangles and a player said 21, her error would be 2. In the *maximum* condition, the team's error equaled the performance of the least accurate player. The team succeeded if the maximum error was less than 2. Finally, in the *minimum* condition, the player with the minimum error determined the team's outcome. The team was successful if at least one player gave the correct solution (i.e. had an error of 0).

The different ways in which the individual contributions combine to determine the team's outcome in our game are representative of the causal structure of many real-world situations (cf. Steiner, 1972). For example, individual contributions often combine additively such as in a game of tug-of-war or the littering of a public place. Sometimes, however, it is the weakest link that determines the group outcome, such as when carrying a heavy object together (e.g. a piano up the stairs). Finally, in other situations, the group is as strong as its best member.

One general knowledge whiz in your group can be enough to secure the pub quiz crown.

In each round of the experiment, participants first performed the counting task and then saw the performance of each player in their team. They were then asked to judge to what extent each player was responsible for the team's outcome. We found that how much responsibility a player received was not only determined by their performance but also sensitive to the group structure. Our task allowed us to test different models of how people might arrive at their responsibility attributions. We found that the structural model (Chockler & Halpern, 2004) predicted participants' attributions significantly better than a simple counterfactual model. Part of the reason for why the simple counterfactual model struggled with predicting people's attributions is that situations of overdetermination occur naturally in our game. For example, if two or more players gave the correct solution in the minimum condition, the team's win was overdetermined. Similarly, if two or more players had an error of 3 or more in the maximum condition, the team's loss was overdetermined. As predicted by the structural model, the responsibility each player received when the outcome was overdetermined was reduced compared to when only one player caused the outcome. However, in contrast to the predictions of the simple counterfactual model, responsibility did not drop to zero once an outcome was overdetermined (cf. Lagnado et al., accepted).

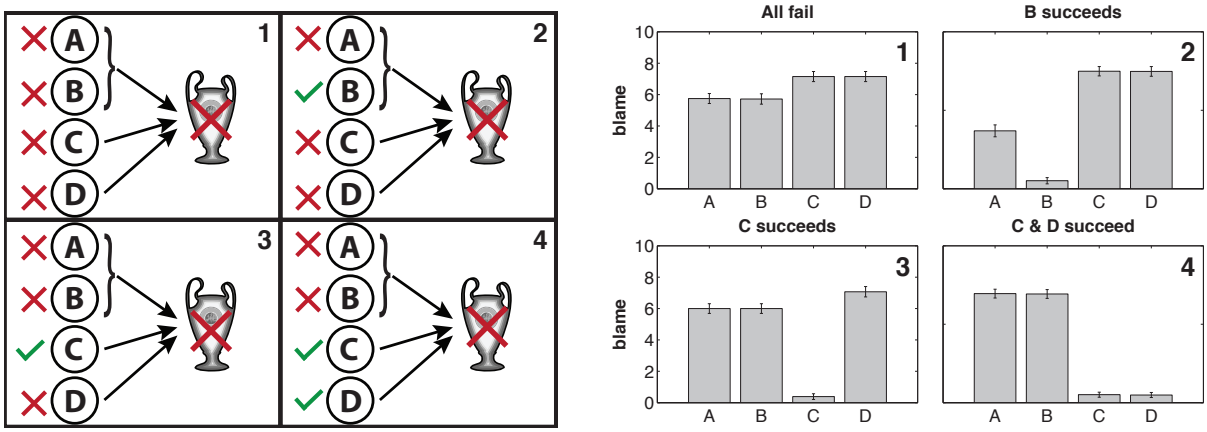
While this experiment established that people's responsibility attributions were sensitive to the causal structure, the experiment's design did not allow us to test some of the more subtle predictions by the structural model. Hence, we ran another series of experiments in which participants acted as external observers. Their task was to evaluate the performance of contestants in a game show. The contestants played a game in which they needed to click on dot on the screen that moved to a random location each time it was clicked. In order to be individually successful in the game, the contestants needed to reach a certain number of clicks within a given time period.

The contestants were then randomly assigned to different teams of four players. In order for the team to win their challenge at least one player out of $A$ and $B$ had to succeed and both players $C$ and $D$ (formally: $\textit{Team outcome} = min(max(A, B), C, D)$). Participants saw four different situations in which a team had lost their challenge (see Figure 7a). For each of the situations, they judged to what extent each player was to blame for the team's outcome (see Figure 7b).

Let us focus on how much blame player $A$ is predicted to receive in the different situations according to the structural model (Chockler & Halpern, 2004).[7] First, note that there is only one situation in which $A$ is pivotal for the team's loss. Namely, the situation in which $A$ and $B$ failed and both $C$ and $D$ succeeded (see Figure 7a, Situation 4). In the four situations that we showed participants, we varied how close $A$ was to being pivotal and tested how this affected the extent to which $A$ was blamed for the team's loss.

In Situation 1 in which all players failed, a minimum of $N = 2$ changes are required to make $A$ pivotal for the loss. Both $C$ and $D$ would need to succeed in order for $A$ to be pivotal. Thus

[7]Chockler and Halpern (2004) distinguish between responsibility and blame, whereby blame is relative to the epistemic state of the agent. In our experiments, we did not vary the agents' epistemic states and thus equate blame with responsibility for negative outcomes (see Robbennolt, 2000; Shaver & Drown, 1986, for theoretical distinctions between the concepts of responsibility and blame).




(a) Different situations

(b) Attributions of blame

Figure 7: Attributions of blame in Experiment 2 by Zultan et al. (2012). The different situations are depicted on the left and participants' average blame attributions to the different players are on the right. Error bars indicate ±1 standard error of the mean. *Note*: The curly brackets in the graphs on the left mean that the contributions of A and B combine in a disjunctive fashion; ✗ = failure, ✓ = success.

the model predicts that $A$'s responsibility is $\frac{1}{3}$. In Situation 2 in which $B$ succeeded, we now need an additional change in order to make $A$ pivotal. We need to change $B$ to having failed and $C$ and $D$ to having succeeded. Because $N = 3$ changes are needed to make $A$ pivotal, $A$'s responsibility is predicted to be $\frac{1}{4}$. In Situation 3 in which $C$ succeeded, only the value of $D$ needs to be changed to make $A$ pivotal and hence $A$'s responsibility is $\frac{1}{2}$. Finally, in Situation 4 in which both $C$ and $D$ succeeded, $A$ is pivotal and thus his responsibility is 1.

As predicted, the results showed that participants' blame attributions were not exclusively determined by a player's performance (i.e. whether they failed or succeeded in their task) but also by the performance of her teammates which influence how close a player was to being pivotal. $A$'s blame for the team's loss reduced significantly when $B$ succeeded (Situation 2 vs. Situation 1). However, in contrast to the model's predictions, $A$'s blame did not increase significantly when only $C$ succeeded (Situation 3 vs. Situation 1). Yet, when both $C$ and $D$ succeeded, $A$ was blamed significantly more than in the baseline condition (Situation 4 vs. Situation 1).

Overall, we see that participants' blame attributions were sensitive to how close a person was to being pivotal. These results are neither predicted by a simple counterfactual model nor a diffusion of responsibility model (Darley & Latané, 1968). According to a simple diffusion model, $A$'s blame should be equal no matter whether either $B$ or $C$ succeeded (Situations 2 and 3). In both of these situations, three players failed their tasks and share the blame for the negative outcome. However, participants' blame attributions show that not only the number of players who failed is important but the causal relationships between the players. Indeed, if $B$ succeeded, $A$ was seen as *less* blameworthy for the negative outcome compared to the baseline condition in which all four players failed.

While participants' blame attributions largely supported the structural model, there were also trends in the data which the model failed to predict. When all players failed, participants attributed significantly more blame to players $C$ and $D$ compared to players $A$ and $B$ (cf.

Figure 7b, Situation 1). However, according to the structural model, each player is equally to blame in this situation. We have seen above that $N = 2$ changes are necessary to render $A$ (or $B$) pivotal (i.e. changing $C$ and $D$). The same is true, however, for player $D$ (or $C$) as well. In order to make $D$ pivotal we need to change $C$ and either $A$ or $B$. Hence, because for both $A$ or $D$ a minimal number of $N = 2$ changes is necessary to render them pivotal, they are predicted to be blamed equally.

However, there is an important asymmetry between players $A$ and $D$. While there is only one way to make $A$ pivotal (via changing $C$ and $D$) there are multiple ways of making $D$ pivotal (e.g. changing $C$ and $A$ or changing $C$ and $B$ or, indeed, changing all $A$, $B$ and $C$). Zultan et al. (2012) took this asymmetry into account and extended Chockler and Halpern's (2004) model. Accordingly, a variable's responsibility is not solely determined by the *minimal* number of changes that are required to render it pivotal but by the number of paths to pivotality. Since there are more paths to make $D$ pivotal than to make $A$ pivotal, $D$ is predicted to receive more responsibility than $A$ (see Zultan et al., 2012, for more details). The novel predictions derived from the extended model were supported in an additional experiment.

Zultan et al. (2012) explained participants' attributions by minimally extending Chockler and Halpern's (2004) model. However, as we will see in the next section, there is an alternative explanation that also accounts for the data. Maybe people's responsibility attributions are not only a function of how close a person was to being pivotal (or of how many paths there were to making a person pivotal) but also influenced by how critical a person's contribution was perceived for a positive team outcome. Before elaborating on this alternative hypothesis, we will briefly discuss work related to the notion of criticality.

# 4 Prospective responsibility

Most work in psychology has focused on on attributions of *retrospective responsibility* – the question of how much an individual is held responsible for an outcome that has already occurred. However, in many situations we have a sense for a person's responsibility even before the actual outcome has occurred (cf. Hart, 2008). For example, we may consider the responsibility that a football manager has for her team's performance or the extent to which a teacher is responsible for his student's educational progress. This future-directed sense of responsibility has been referred to as *role responsibility* (Hart, 2008) or *prospective responsibility* (Cane, 2002; Lagnado et al., accepted).[8]

Consider the following scenario (cf. Zultan et al., 2012): four participants in a cooking show are randomly assigned to one of three different roles. Two will separately prepare starters, one will prepare the main and one the dessert. Each dish is evaluated by judges and can either pass or fail. In order for the team to be successful, they have to make sure that at least one starter passes and that both the main and the dessert pass. Hence, the structure of the task is identical to the one discussed above (see Figure 7). Given this setup, whose contribution is more critical for the team's outcome? Adam who prepares one of the two starters (only one of which is necessary for the team's success) or Claire who prepares the main? If you feel that Claire is

[8]We will use the terms *criticality*, *prospective responsibility* and *(in-)dispensability* interchangeably.

more critical, you agree with what most participants in our experiment said (see below).

There has been relatively little work on how people attribute prospective responsibility. Hamilton (1978) argued that (retrospective) attributions of responsibility are not only influenced by what someone did but also by expectations about what the person should have done (see also Schlenker, Britt, Pennington, Murphy, & Doherty, 1994). More precisely, a person's responsibility is affected by their role: while a policeman is clearly responsible if he does not intervene when a pedestrian is mugged, a fireman would be held responsible to a lesser degree (if at all). In the cooking show, one could argue that Claire has a more central role than Adam and that people will expect her to put in more effort. However, even if this were the case, we would still need to explain where these expectations come from.

In a related domain, research on efficacy has supported the positive relationship between perceived outcome control and exertion of effort (Bandura, 1977). A person exhibits efficacy to the extent that she can bring about the changes in the world she desires to obtain. Individual efficacy is both influenced by the person's ability and the difficulty of the attempted task. In a group context, whether a person is efficacious in achieving her desired outcome also depends on the behavior of her group members.

Kerr and Bruun (1983) have shown that the extent to which a person perceives her individual efforts to be dispensable for the group's success influences their performance. In a series of experiments, they varied three factors that were predicted to have an effect on an individual's perceived dispensability and exertion of effort: group size, task structure and member ability. Consider the cooking show example once more in which the preparation of the starters follows a disjunctive task structure. In what situation would you predict Adam (who prepares one of the starters) to feel more critical for the team's success: if his partner is a very good chef or if his partner's cooking skills are rather poor? Kerr and Bruun (1983) found that participants exerted more effort and perceived their contribution as more important when their ability was higher than the ability of their teammate for a task with a disjunctive structure. In contrast, if the task followed a conjunctive structure and the weakest member determined the group's overall performance, participants exerted more effort and perceived their contributions to be more critical when their level of ability was low compared to their group member.

More recently, research on social dilemmas has looked more closely at efficacy (also referred to as *criticality*) in groups (Au, Chen, & Komorita, 1998; Yu, Au, & Chan, 2009). The step-level public goods game is a well-studied experimental setup of a social dilemma situation (Rapoport, 1987). Here is an example of how the paradigm works: Consider a group of five players each of whom receives an initial endowment of \$5. Each of them can decide individually whether they want to keep their money or contribute it toward the public good. The public good is provided if at least three of the group members contribute their endowment. If the good is provided, each group member receives an additional \$10 no matter whether they themselves contributed their endowment or not.

What would you do if you were a participant in this experiment? Would you keep the money or contribute it to the common pot? If you kept the money and at least three of your group members contributed theirs you would end up with \$15. However, if not enough people contribute, then the public good will not be provided. If you contribute the money, you increase

the chances that the public good will be provided. However, there is also a chance that you'll go away with no money (in case the public good is not provided) or that your contribution turns out to be superfluous (if, for example, three of the other group members also decide to contribute their money).

Rapoport (1987) has provided a definition according to which a person's criticality is given by the probability that their contribution will make a difference to the outcome. This probability depends on the size of the group, the provision point (i.e. the minimum number of players whose contributions is necessary for the public good to be provided), and the probability that the other group members will contribute.[9] Subsequent research has shown that perceived criticality correlates positively with the likelihood to contribute (see, e.g. De Cremer & Dijk, 2002; Kerr, 1996).

### 4.1 A simple model of criticality

Most research on criticality has focused on symmetric task structures. However, as illustrated with the cooking show example above, there are situations in which the task structure is asymmetric. In such situations, the individual's contributions affect the probability that a successful outcome will come about in unequal measures (cf. Kerr, 1992). In the cooking show, Claire's main has to get a pass from the judges in order for the team to be able to succeed. Adam's starter, in contrast, is not a necessary condition for the team's success. Even if his dish does not pass, the team can still succeed in case the other starter passes. How do people judge the importance of each person's contribution for the team's successful outcome in such situations? Here, we propose a very simple model. We predict that people first consider how many subtasks are necessary in order for the team to succeed. Each necessary subtask is viewed as fully critical and thus receives a criticality value of 1. If subtasks are shared between multiple players in a disjunctive fashion, the criticality for the subtask is divided between the players.

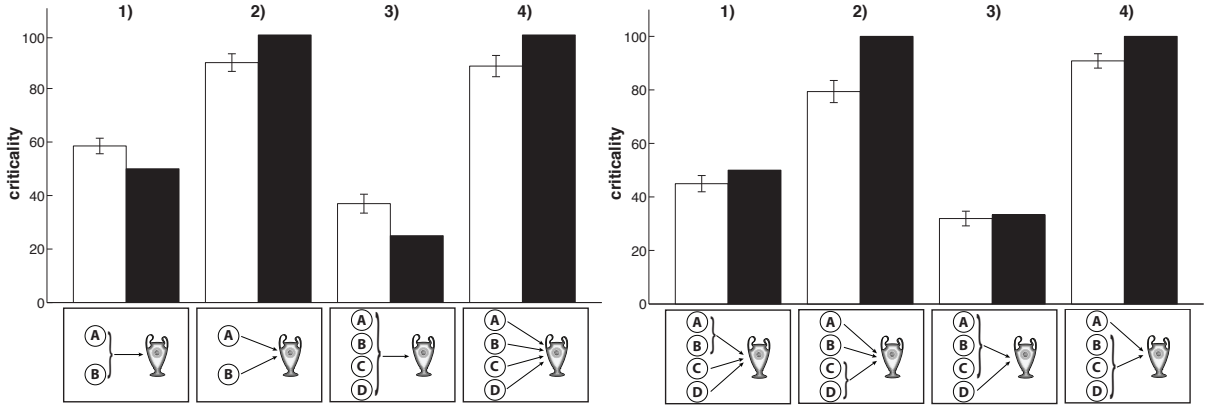
Applied to our cooking show example, the model makes the following predictions. First, each of the necessary subtasks (starter, main and dessert) are viewed as fully critical. Since the main and dessert are prepared by only one person each, both are fully critical. The subtask of the starter, however, is shared in a disjunctive manner. Only one of the starters needs to pass in order for the subtask to be successful. Thus, our model predicts that each of the two chefs doing starters receives a criticality of $\frac{1}{2}$.

In order to test our model, we again employed the dot-clicking game to model different team challenges of interest (Lagnado et al., accepted; Zultan et al., 2012). This time, participants viewed four different challenges on the screen simultaneously and were asked to judge how critical Player $A$ was for the team's outcome in each challenge.

Figure 8 shows participants' ratings in white together with the predictions of our simple criticality model in black. The two sets of challenges in Figure 8a and 8b were presented separately. The criticality model predicts participants' ratings very well with $r = .97$ and $RMSE = 11.15$. Player $A$ is rated highly critical when his performance is necessary for the

[9] Research in voting theory has derived similar definitions (cf. Banzhaf, 1964; Felsenthal & Machover, 2004; Shapley & Shubik, 1954). In voting theory, one is concerned with the power that an individual person or state has to influence the outcome of an election. Different indices have been developed that link a person's voting power with the probability that their vote will be pivotal.

team's success (cf. Challenges 2 and 4 in Figure 8a and Figure 8b). In situations in which $A$ shares the subtask with other players in a disjunctive way, his criticality reduces with the number of people who share the task (cf. Challenges 1 and 3 in Figure 8a and Figure 8b).




(a) First set of challenges

(b) Second set of challenges

Figure 8: Mean criticality judgments for player $A$ (white) and model predictions (black) for two different sets of challenges. Error bars indicate $\pm 1$ standard error of the mean.

## 5 Retrospective responsibility (revisited)

In the previous section, we have seen that some accounts predict that the extent to which a person is held responsible for an outcome is not only dependent on their action but also on expectations attached to the person's role in a given situation (Hamilton, 1978; Schlenker et al., 1994). Our simple criticality model captures the different degrees to which individuals' contributions are necessary for the outcome.[10] Forsyth, Zyzniewski, and Giammanco (2002) have argued that responsibility often does not diffuse equally between the members of a group (cf. Darley & Latané, 1968) but rather "unevenly... with more being apportioned to group members who occupy more central positions in the group" (Forsyth et al., 2002, p. 55). In their discussion, they hypothesize that responsibility allocations are likely to be influenced by the way in which individual contributions combine to yield the group outcome. In terms of our framework, the different ways in which the individual contributions combine affects the extent to which they are perceived to be critical for the outcome which in turn influences how much responsibility is attributed to them.

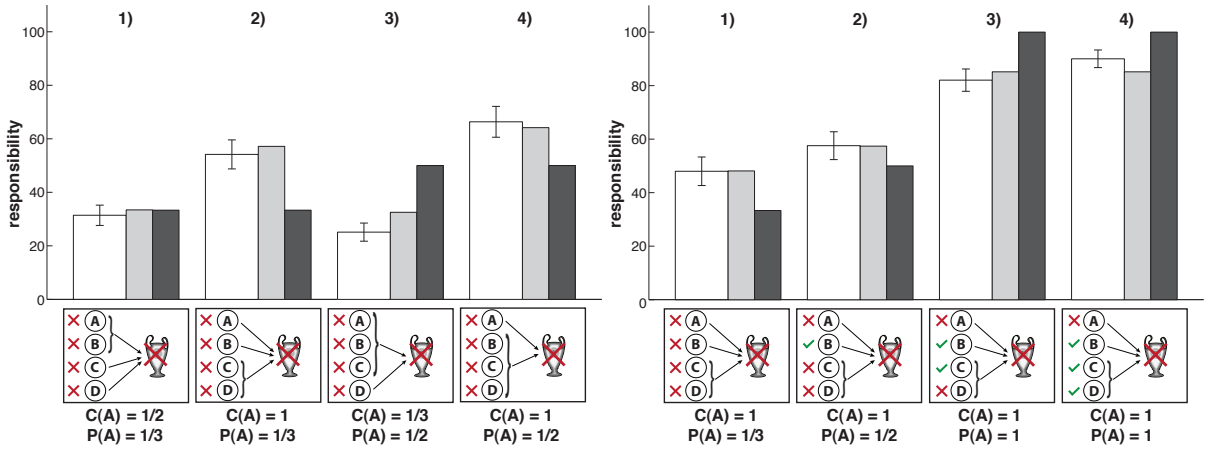
Recently, we devised an experiment to test to what degree attributions of responsibility are influenced by (i) *pivotality* and (ii) *criticality* (Lagnado et al., accepted). With pivotality, we mean how close a person's contribution was to making a difference to the outcome (cf. Chockler & Halpern, 2004; Zultan et al., 2012). As outlined above, we can manipulate a group

[10]Of course, Hamilton's (1978) conception of *role responsibility* is considerably richer than what we capture in our simple team setup. However, we believe that these abstract tasks are a good starting point for looking at how different roles affect attributions of responsibility. Indeed, our broad understanding of role responsibility appears to be closely related to how Hart (2008) understood the concept: "If two friends, out on a mountaineering expedition, agree that one shall look after the food and the other the maps, then the one is correctly said to be responsible for the food, and the other for the maps, and I would classify this as a case of role-responsibility" (Hart, 2008, p. 212).

member's pivotality by varying the performance of the other team members (cf. Figure 7). We manipulated the criticality of a person by varying the task structures (cf. Figure 8). In line with Forsyth et al. (2002), we hypothesized that participants' responsibility attributions would not only be affected by how close a person was to making a difference to the outcome (i.e. their pivotality) but also by how important a person's contribution was perceived to be for the team's result (i.e. their criticality). Thus, a model that predicts attributions of responsibility as a weighted function of a person's criticality and pivotality should perform better than a model which only uses pivotality or criticality as a predictor.

In our experiment, we used the eight different team structures shown in Figure 8. Again, participants viewed four different situations simultaneously on the screen. This time, however, the performance of each player (i.e. whether they had succeeded or failed) as well as the group outcome (i.e. whether the team had won or lost) were shown. Participants judged how responsible Player A was for the team's result in the different situations.

Participants viewed nine different sets of situations in total. In these sets, we systematically varied player $A$'s pivotality and criticality. Figure 9 shows two of the sets of situations that participants saw.[11] In the four situations shown in Figure 9a, all players failed in their individual task. Note that the task structures differed between the four situations. The task structures affect how critical player $A$ is and also how close player $A$ was to being pivotal in the different situations.




(a) One set of situations

(b) Another set of situations

Figure 9: Mean responsibility attributions to player $A$ (white bars) and model predictions (criticality-pivotality model = gray, pivotality model = black) for two different sets of challenges. C(A) = A's criticality, P(A) = A's pivotality, ✗ = failure, ✓ = success. Error bars indicate ±1 standard error of the mean.

Let us consider Situation 1 in Figure 9a first. The task structure in this situation is the same as the cooking show example mentioned above. Our simple criticality model predicts that player $A$ will be perceived to be $\frac{1}{2}$ critical for the team's result in this challenge. As it turned out, all players happened to fail their individual dot-clicking game and the team thus lost the challenge. The minimal number of changes that are required to render player $A$ pivotal in this situation equals $N = 2$ and thus $A$'s pivotality is $P(A) = \frac{1}{3}$. In Situation 2 in Figure 9a, $A$'s

[11] See Lagnado et al. (accepted) for the full set of data.

criticality according to our model is $C(A) = 1$. Again, a minimum of two changes is required to render $A$ pivotal and his pivotality is thus $P(A) = \frac{1}{3}$.

Thus, while player $A$'s pivotality in Situation 1 and 2 is the same, she is more critical in Situation 2 than in Situation 1. Participants assigned player $A$ more responsibility for the team's loss in Situation 2 than in Situation 1. Similarly, the pivotality of player $A$ is identical in Situation 3 and Situation 4 ($P(A) = \frac{1}{2}$). However, while $A$'s criticality is only $C(A) = \frac{1}{3}$ in Situation 3, he is fully critical in Situation 4. Participants attributed more responsibility to player $A$ in Situation 4 than in Situation 3. Note also that despite the fact that only one change is necessary to make $A$ pivotal in Situation 3 and two changes would be needed to make $A$ pivotal in Situation 1, $A$ is blamed significantly *more* in Situation 1 than in Situation 3.

While these results show that participants' responsibility attributions cannot be explained in terms of pivotality only, they do not yet give strong support for our hypothesis that both criticality and pivotality influence people's attributions of responsibility. A more parsimonious explanation for this pattern of results would be that participants assign responsibility based on criticality only. However, our experiment also included situations that rule out this responsibility-equals-criticality explanation.

Figure 9b shows a set of situations in which the task structure was identical in all four situations. Keeping $A$'s criticality constant ($C(A) = 1$), we varied $A$'s pivotality via the performance of the other group members. In Situation 1, all players failed and $A$'s pivotality is $P(A) = \frac{1}{3}$. In Situation 2, player $B$ succeeded and hence $A$'s pivotality increases to $P(A) = \frac{1}{2}$, because only one more change would be necessary to render $A$ pivotal. Finally, in Situations 3 (in which both player $B$ and $C$ succeed) and 4 (in which all players except $A$ succeed), $A$'s pivotality is $P(A) = 1$. Participants' attributions of responsibility increased with pivotality when criticality was held constant (in line with the results of the experiment by Zultan et al., 2012, reported above, cf. Figure 7).

Overall, the results show that participants' responsibility attributions are sensitive both to how close a person's contribution was to making a difference to the outcome (i.e. their pivotality) and the extent to which their contribution was perceived to be critical for the team's result. A model that predicts participants' attributions in terms of both pivotality and criticality explains the data better ($r = .97, RMSE = 6.34$) than a model that uses just pivotality ($r = .77, RMSE = 21.05$) or just criticality ($r = .67, RMSE = 25.81$) as a predictor (see Lagnado et al., accepted, for a more detailed analysis of the results).

## 6 Conclusion

How do people attribute responsibility in situations in which the contributions of multiple people combine to yield a group outcome? In this paper, we have argued that attributions of responsibility are not solely determined by what actually happened but also influenced by considerations about other possible worlds that could have come about. We have discussed empirical evidence that attributions of responsibility are sensitive to whether a person's contribution made a difference to the outcome. While *making a difference* and *being responsible* appear to go hand in hand in many situations, we have also seen that there are situations in

which a person's action does not make a difference to the outcome but we nevertheless feel that the person is responsible to some degree (Gerstenberg & Lagnado, 2010, 2012; Lagnado et al., accepted; Zultan et al., 2012). None of the prevalent models in the psychological literature on responsibility attribution can account for these effects (Brewer, 1977; Fincham & Jaspars, 1983; Petrocelli et al., 2011; Spellman, 1997).

While some have argued against a close relationship between attributions of causality (or responsibility) and counterfactuals (Mandel, 2003; Sartorio, 2004), we have shown that a richer conception of counterfactuals resolves some of the problems identified in the literature. In particular, we have demonstrated how to resolve the problem of assigning responsibility in situations of causal overdetermination. While a person is generally held responsible if she made a difference to the outcome, she can *also* be held responsible if what she did made no difference in the actual situation *but* would have made a difference in another possible situation. Our framework for understanding responsibility attributions in terms of counterfactuals defined over causal models provides the conceptual tools to address the problem of assigning responsibility in complex situations with multiple causes. We can derive quantitative predictions from this framework about how we expect responsibility to be distributed and test these predictions in simple experiments such as the ones outlined in this chapter. The results of these experiments have shown that responsibility attributions are not only affected by how close a person's contribution was to making a difference to the outcome but also by how critical their contribution was for a positive outcome in the first place.

Our framework defines attributions of responsibility in terms of counterfactual dependence under possible contingencies as implied by the causal structure of the situation. It thus operates on a high level of abstraction. However, it is clear that responsibility attributions are not just sensitive to whether a certain event was present or absent – it matters *how* the event came about (cf. Lombrozo, 2010). So far, our framework does not distinguish between situations in which the variables in the causal network represent the performances of individuals in a sports context, strategic decisions or physical events. This makes our approach very general and applicable to many situations. In future research, we will investigate how additional factors, such as intentions or norms influence responsibility attributions.

## References


Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

Au, W. T., Chen, X. P., & Komorita, S. S. (1998). A probabilistic model of criticality in a sequential public good dilemma. *Organizational Behavior and Human Decision Processes*, *75*(3), 274–293.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Banzhaf, J. F. (1964). Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, *19*, 317–343.

Brewer, M. B. (1977). An information-processing approach to attribution of responsibility. *Journal of Experimental Social Psychology*, *13*(1), 58–69.

Campbell, J. (2008). Interventionism, control variables and causation in the qualitative world. *Philosophical Issues*, *18*(1), 426–445.

Cane, P. (2002). *Responsibility in law and morality.* Oxford: Hart Publishing.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Chockler, H., Halpern, J. Y., & Kupferman, O. (2008). What causes a system to satisfy a specification? *ACM Transactions on Computational Logic*, *9*(3), 20.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*(1), 525–556.

De Cremer, D., & Dijk, E. van. (2002). Perceived criticality and contributions in public good dilemmas: A matter of feeling responsible to all? *Group Processes & Intergroup Relations*, *5*(4), 319–332.

Felsenthal, D., & Machover, M. (2004). A priori voting power: what is it all about? *Political Studies Review*, *2*(1), 1–23.

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of responsibility: From man the scientist to man as lawyer. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13, p. 81). New York: Academic Press.

Fincham, F. D., & Jaspars, J. M. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology*, *22*(2), 145–161.

Forsyth, D. R., Zyzniewski, L. E., & Giammanco, C. A. (2002). Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, *28*(1), 54–65.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, *66*(23), 829–839.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (submitted). Back on track: Backtracking in counterfactual reasoning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin, TX:

Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

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

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review*, *19*(4), 729–736.

Gregory, C. O. (1932). Vicarious responsibility and contributory negligence. *The Yale Law Journal*, *41*(6), 831–852.

Halpern, J. Y., & Hitchcock, C. (forthcoming). Graded causation and defaults.

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.

Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology*, *41*(4), 316–328.

Hart, H. L. A. (2008). *Punishment and responsibility*. Oxford: Oxford University Press.

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. Oxford University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc.

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.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*(6), 273–299.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *11*, 587–612.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128.

Kerr, N. L. (1992). Efficacy as a causal and moderating variable in social dilemmas. In *Social dilemmas: Theoretical issues and research findings* (pp. 59–80).

Kerr, N. L. (1996). "Does my contribution really matter?": Efficacy in social dilemmas. *European Review of Social Psychology*, *7*(1), 209–240.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*(1), 78–94.

Kerr, N. L., & Kaufman-Gilliland, C. M. (1997). "... and besides, I probably couldn't have made a difference anyway": Justification of social dilemma defection via perceived self-inefficacy. *Journal of Experimental Social Psychology*, *33*(3), 211–230.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4),

315–365.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (accepted). Causal responsibility and counterfactuals. *Cognitive Science*.

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343–356.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, *132*(3), 419–434.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

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.

McCoy, J., Ullman, T., Stuhlmüller, A., Gerstenberg, T., & Tenenbaum, J. B. (2012). Why blame Bob? Probabilistic generative models, counterfactual reasoning, and blame attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1996–2001). Austin, TX: Cognitive Science Society.

Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology*, *59*(6), 1111–1118.

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

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, *331*(6023), 1401–143.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*(4), 663–685.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

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.

Rapoport, A. (1987). Research paradigms and expected utility models for the provision of step-level public goods. *Psychological Review*, *94*(1), 74–83.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61.

Robbennolt, J. K. (2000). Outcome severity and judgments of "responsibility": A meta-analytic review. *Journal of Applied Social Psychology*, *30*(12), 2575–2609.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–148.

Sartorio, C. (2004). How to be responsible for something without causing it. *Philosophical Perspectives*, *18*(1), 315–336.

Schaffer, J. (2003). Overdetermining causes. *Philosophical Studies*, *114*(1), 23–45.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, *101*(4), 632–652.

Shapley, L., & Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, *48*(3), 787–792.

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

Shaver, K. G., & Drown, D. (1986, April). On causality, responsibility, and self-blame: a theoretical note. *Journal of Personality and Social Psychology*, *50*(4), 697–702.

Shultz, T. R., Jaggi, C., & Schleifer, M. (1987). Assigning vicarious responsibility. *European Journal of Social Psychology*, *17*(3), 377–380.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348.

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*(4), 241–264.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search.* The MIT Press.

Steiner, I. D. (1972). *Group process and productivity.* Academic Press.

Strawson, P. F. (2008). *Freedom and resentment and other essays.* Taylor & Francis.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49–54.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *56*(2), 161–169.

Widerker, D., & McKenna, M. (2006). *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities.* Ashgate Publishing Company.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford University Press, USA.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Yablo, S. (2002). De facto dependence. *The Journal of Philosophy*, *99*(3), 130–148.

Yu, C., Au, W., & Chan, K. K. (2009). Efficacy = endowment × efficiency: Revisiting efficacy and endowment effects in a public goods dilemma. *Journal of Personality and Social Psychology*, *96*(1), 155–169.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, *125*(3), 429–440.