Making a Difference

Responsibility, Causality and Counterfactuals

Tobias Gerstenberg

Cognitive, Perceptual and Brain Sciences Research Department

University College London

Thesis submitted for the degree of

Doctor in Philosophy (PhD)

January, 2013

Abstract

In this thesis, I develop a general framework of how people attribute responsibility. In this framework, people's responsibility attributions are modelled in terms of counterfactuals defined over a causal representation of the situation. A person is predicted to be held responsible to the extent that their action made a difference to the outcome. Accordingly, when attributing responsibility we compare what actually happened with the outcome in a simulated counterfactual world in which the person's action had been different. However, a person can still be held responsible for an outcome even if their action made no difference in the actual situation. Responsibility attributions are sensitive to whether a person's action would have made a difference in similar counterfactual situations. Generally, responsibility decreases with the number of events that would have needed to change from the actual situation in order to generate a counterfactual situation in which the person's action would have been pivotal. In addition to how close a person was to being pivotal, responsibility attributions are influenced by how critical a person's action was perceived prior to the outcome.

The predictions derived from this general framework are tested in a series of experiments that manipulate a person's criticality and pivotality by varying the causal structure of the situation and the person's mental states. The results show that responsibility between the members of a group diffuses according to the causal structure which determines how individual contributions combine to yield a joint outcome. Differences in the group members' mental states, such as their knowledge about the situation, their expectations about each other's performance as well as their intentions, also affect attributions. Finally, I demonstrate how this general framework can be extended to model attributions for domains in which people have rich, intuitive theories that go beyond what can be expressed with simple causal models.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

Signature: London, January 21, 2013

(Tobias Gerstenberg)

To my parents

Acknowledgements

> Life [as a PhD student] is full of misery, loneliness, and suffering – and it's all over much too soon.
>
> – **Woody Allen**

> I like to praise and reward loudly, to blame quietly.
>
> – **Catherine II**

It looks like my life as a PhD student is finally coming to an end. Woody Allen is right – it's all over much too soon (despite the occasional misery and loneliness). The title of my thesis is 'Making a Difference'. First, I thought, that sounds a little bit over the top. Since it's commonly assumed that "nobody reads your thesis", this piece of work will probably not make a difference to too many people (apart from my examiners who have to wade their way through it). However, because I felt the title fits so well with the theme of responsibility attribution explored in this thesis, I could not resist. Here, I would like to follow Catherine II and praise loudly some of the many people who have made a difference to my life.

I consider myself incredibly lucky to have been supervised by an inspiring scientist, a wonderful storyteller and a groovy bass player who toured with the likes of Jerry Lee Lewis and Van Morrison. A day in the life of Dave Lagnado's PhD student might start with a heated debate about who is to blame for Chelsea's loss and end in a bar in Soho, sipping red wine and listening to hours of blues and jazz by 'Tracy Ray and the Mojos' – with Dave on double bass. Thanks Dave, for having made the last couple of years such an exceptional experience!

I would also like to thank Nick Chater for his support and encouragement especially at the beginning of my PhD. Throughout my PhD, I have had the fortune to interact and work with many brilliant minds. This thesis is the product of having collaborated with Catherine Cheung, Anastasia Ejova, Noah Goodman, Yaakov Kareev, David Lagnado, John McCoy, Simeon Schächtele, Maarten Speekenbrink, Andreas Stuhlmüller, Josh Tenenbaum, Tomer Ullman and Ro'i Zultan. A big thanks to all of you!

Extra thanks go to Noah Goodman and the members of the Computation & Cognition Lab at Stanford as well as to Josh Tenenbaum and the members of the Computational Cognitive Science Group at MIT for having hosted me during my three-month visit to the US last year.

Thanks to Stephanie Baines, Christos Bechlivanidis, Neil Bramley, Caren Frosch, Adam Harris, Anne Hsu, Petter Johansson, Irma Kurniawan, Jens Koed Madsen, Milena Nikolic, Chris Olivola, Fiona Patrick, Ramsey Raafat, Stian Reimers, Costi Rezlescu, Adam Sanborn, Katya Tentori, Zoe Theocharis, Konstantinos Tsetsos, Marion Vorms and Erica Yu for having made UCL a great place to work and London a wonderful place to live.

Thanks also to David Shanks and the Shanks' lab members Tom Beesley, Chris Berry, Tom Hardwicke, Manos Konstantinidis, Rosalind Potts, Sarah Smith, Maarten Speekenbrink, Miguel Vadillo and Emma Ward for enduring some of my presentations at the lab meetings.

Certainly one of the best aspects about doing a PhD is getting in touch with so many interesting people. Thanks to Denis Hilton, Richard Holton, Joshua Knobe, William Jiménez-Leal, Björn Meder, Jonas Nagel, Magda Osman, Philip Pärnamets, Anne Schlottmann, Eric Schulz, Shaul Shalvi, Steve Sloman, Michael Waldmann and Alexander Wiegmann for valuable feedback and good times.

"A friend is someone who knows all about you and still loves you." (Elbert Hubbard) Thanks to my friends in London, Germany and elsewhere for knowing *and* loving me. Thanks also to four eras of housemates who have helped to make London feel like home. Extra thanks to Lisa – my English mother!

I would like to express my gratitude to the AXA Research Fund for having provided financial support throughout my PhD.

Writing this thesis has been made considerably more fun thanks to LaTeX, Grooveshark and halloumi wraps from Restaurant Leziz combined with Larry David's 'Curb your Enthusiasm' for lunch.

Finally, I would like to thank my parents Bodo and Rosel – you certainly made the biggest difference to my life –, my brother Timo and my sister Miriam for all their love and support.

Published and submitted articles

The chapters in this thesis are based on the following articles:

Chapter 4: Causal Structure and Responsibility

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. (accepted). Attributing responsibility: Actual and counterfactual worlds. In J. Knobe, T. Lombrozo, & S. Nichols (Eds.), *Oxford Studies of Experimental Philosophy*.

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

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

Chapter 5: Mental States and Responsibility

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. (pp. 720–725). Austin, TX: Cognitive Science Society.

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

Gerstenberg, T., Lagnado, D. A., & Kareev, Y. (2010). The dice are cast: The role of intended versus actual contributions in responsibility attribution. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1697–1702). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Lagnado, D. A., Speekenbrink, M., & Cheung, C. (2011). Rational order effects in responsibility attributions. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. (pp. 1715–1720). Austin, TX: Cognitive Science Society.

Schächtele, S., Gerstenberg, T., & Lagnado, D. A. (2011). Beyond outcomes: The influence of intentions and deception. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin, TX, 2011 (pp. 1860–1865). Cognitive Science Society.

Chapter 6: Beyond Bayes Nets

Gerstenberg, T. & Goodman, N. D. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). 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.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> History is the story of events, with praise or blame.
>
> – **Cotton Mather**

"McDonald, 27, tied with Olivia Ballou for a seat on Walton City Council in Kentucky on 669 votes, The Kentucky Enquirer reported. When his wife Katie rang him with 10 minutes to go before the polls closed to say she hadn't had time to vote, he told her not to bother. 'If she had just been able to get in to vote, we wouldn't be going through any of this,' McDonald said. 'You never think it will come down to one vote, but I'm here to tell you that it does.'

McDonald, whose fate is likely to be decided by a toss of a coin if a recount does not split them, did not blame his wife, who works nights as a patient care assistant at a hospital and is finishing training as a nurse. 'She feels bad enough', he said. 'She worked extra hours, goes to school and we have three kids, so I don't blame her. She woke up about ten minutes before the polls closed and asked if she should run up, but I told her I didn't think one vote would matter.'"[1]

McDonald learned the hard way that sometimes 'every vote counts'. If only his wife had managed to go and vote, he would have won the elections (assuming she would have voted for him). From the perspective of a purely self-interested *homo economicus* who is just interested in maximising his or her individual utility, the fact that people vote is paradoxical (Downs, 1957): the individual costs for voting (such as waiting for hours in long queues) normally outweigh the often minuscule chances of casting the pivotal vote – the vote that would make a difference to the outcome of the election. However, the fact that a large proportion of people are indeed willing to bear the costs of voting suggests that they are not only concerned about whether or not their vote will make a

[1] Retrieved on 18/11/2012 from: `http://www.heraldsun.com.au/news/world/robert-bobby-mcdonald-tied-with-olivia-ballou-on-walton-city-council-after-he-told-wife-not-to-vote/story-fnd134gw-1226513833424`

difference. With a different conception of rationality in mind, one that grants people a sense of social responsibility if they regard their own vote as a potential contribution to a general good, it can indeed be rational to vote (De Cremer & van Dijk, 2002; Edlin, Gelman, & Kaplan, 2007).

This thesis will not be concerned with analysing what motivates people to vote but rather with the related question of how people attribute responsibility to individuals for outcomes which resulted from the actions of several people. Although McDonald had told his wife that one vote won't matter and consequently did not blame her, we might still wonder whether she nevertheless feels some responsibility for the outcome. It is also conceivable that McDonald's reaction would have been quite different if he had indeed told his wife to go and vote.

Central to this thesis is the argument that attributions of responsibility are closely linked with how much a person's contribution was perceived to have made a difference to the outcome. One way of assessing the extent to which a person's contribution made a difference is via considering what would have happened if the person's action had been different. That is, we compare the actual world with another possible world in which we alter the person's action and evaluate whether the outcome in this counterfactual world would have changed. If we are confident that the same outcome would have prevailed not matter whether or not a person had acted, the intuition is strong that the person's responsibility for the outcome is at least reduced. If, for example, McDonald had found himself in a situation in which he had lost by two votes, the outcome of the election would have been the same whether or not his wife had voted. However, although her action would not have made a difference in this situation, it would have made a difference in another possible situation in which her husband had only lost by one vote instead of two.

A central finding of this thesis is that people's attributions of responsibility are not only determined 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 that could have come about. All else being equal, a person's responsibility decreases the further away the actual world is from a world in which the person's contribution would have made a difference. Hence, McDonald's wife should feel less responsible when her husband lost by ten votes rather than one vote.

We will also see that people's responsibility attributions are not only sensitive to how close a person's contribution was to making a difference after the outcome but also to how important their contribution was perceived *before* the outcome had been realised. For example, imagine that McDonald was a Mormon and had five wives. Let us assume further that McDonald (and his wives) happen to know that if at least one of his wives votes for him, he will certainly win the election. However, all of his wives happen to be exceptionally busy on the day and none of them manages to go and vote. As it turns out, he lost the election by one vote. While each of his wives would have

made a difference in the actual situation, it nevertheless feels that their responsibility is somewhat reduced (compared to the one wife scenario) due to the fact that their actions could have compensated for each other. The vote of each individual wife in the Mormon scenario is less critical for the outcome compared to a situation in which McDonald had only one wife (who knew that her vote would be necessary).

While the structure underlying an election is fairly straightforward (especially for elections with a simple majority rule), there are many situations in which the contributions of several people combine in a more complex manner to determine the joint outcome. We might wonder, for example, how much responsibility an individual player in a team sport, such as football, carries for the team outcome. As another core theme of this thesis, we will see that people's responsibility attributions are sensitive to the underlying causal structure of the situation. In order to evaluate whether the outcome would have been different but for a person's action, we need to have a good understanding of the causal structure of the situation which dictates what would have happened under different possible contingencies.

In this thesis, I propose a general framework that conceptualises attributions of responsibility in terms of counterfactuals defined over a causal representation of the situation. Accordingly, when attributing responsibility, people compare what actually happened with what they think would have happened if the causal event of interest had been different. People's causal representation of the situation determines whether the considered counterfactual would have undone the outcome. Furthermore, the causal representation also influences what other contingencies are considered in which the causal event of interest could have made a difference. Responsibility attributions are not solely determined by what actually happened but also by what could have happened in other contingencies that are (causally) similar to the actual world.

The structure of the thesis is as follows:

In Chapter 2, I will provide an overview of the key frameworks and findings in attribution theory. The focus will be on theories of responsibility attribution. Previous research has already argued for a close relationship between attributions of responsibility, causality and counterfactuals. However, I will show that these theories have not succeeded in making the relationships between these concepts sufficiently precise.

In Chapter 3, I will show, based on recent formal work in the cognitive sciences – most notably the work of Judea Pearl (2000) – how the relationship between causality, counterfactuals and responsibility can be made more precise via a causal Bayes net framework. Causal Bayes nets not only support claims about general causal relationships but also about whether one event was the actual cause of another event in a particular situation. I will show how from this notion of *actual causation*, we can derive a model which predicts attributions of responsibility in terms of counterfactuals defined over the causal structure of a situation (Chockler & Halpern, 2004).

Equipped with the necessary theoretical background and the formal tools for expressing responsibility in terms of counterfactuals over causal models, Chapter 4 will demonstrate empirically that people's responsibility attributions are sensitive to the underlying causal structure which specifies the ways in which multiple causes combine to generate the outcome. Focussing on responsibility attributions in group settings, we will see that the causal relationships between the agents in a group determine how the collective responsibility for the outcome diffuses between the group members.

While Chapter 4 focuses on how the causal structure influences responsibility attributions, Chapter 5 will zoom into the heads of the individuals within the group. I will discuss a series of experiments that investigate how manipulations of the group members' mental states systematically affect attributions of responsibility. We will see that it not only matters whether an individual's attribution made a difference to the outcome but also whether the person *knew* that their contribution was critical for the group's success. Furthermore, I will show that attributions of responsibility are sensitive to the team members' actual performance as well as their underlying skill level. Finally, I will discuss a series of experiments in which we investigated how people attribute responsibility in noisy environments in which intended contributions and actual contributions can come apart.

In Chapter 6, I will argue based on evidence from two different experimental domains, that it is sometimes necessary to go beyond the representational capabilities of simple Bayesian networks in order to adequately model people's attributions. In the first part, I will show how people's intuitive understanding of physics informs their causal attributions. In support of the general framework, participants' attributions are well explained by assuming that they compare what actually happened with the outcome of a mental simulation about what would have happened if the causal event of interest had not taken place. In the second part, we will see how people make inferences about an agent's disposition based on diverse patterns of evidence in a way that is difficult to capture with simple Bayesian networks. However, when we assume that people's mental representations are structured compositionally, we can begin to develop models that show how people's inferences are possible.

Finally, Chapter 7 will wrap things up, summarise the main lessons learned and point out avenues for future research worth pursuing.

Bernard Weiner (1995), whose work will be discussed in Chapter 2, said that "the assignment of responsibility and its consequent affects and behaviors are not confined to the courtroom, for life itself is a courtroom where we all act as judges." (p. 23) Figure 1.1 provides a glimpse into the courtroom of this thesis, whereby the size of the different words corresponds to the frequency with which they appear. The word cloud also provides a concise representation of everything this thesis is *not* about. For example, the fact that the word 'moral' did not make it into the top 50 of the mostly used words,

Figure 1.1: A wordle based on the content of the thesis (see `http://www.wordle.net`).

illustrates that this thesis won't be concerned with the attribution of *moral responsibility* in particular. Rather, it has at its focus a more general notion of responsibility that is closely linked with causality and counterfactuals (and these terms do indeed appear in the word cloud). Metaphorically speaking, a great deal of this thesis will be concerned with turning this word cloud into a causal model showing how the different concepts are interrelated: A **causal model** of **participants**' **responsibility attributions** to **players** in a **team** – would the **counterfactual outcome** of the **game** have been **different** if the **individual player**'s **performance** had been **different**?

Chapter 2

Theoretical Frameworks of Attribution

> The causes of events always interest us more than the events themselves.
>
> – **Cicero**

> Happy is he who has been able to perceive the causes of things.
>
> – **Virgil**

As the quotes by Cicero and Virgil suggest, people have an intrinsic motivation to understand how the world works. We are curious when we find a bouquet of flowers on our door step on a Sunday morning and happy to find out that they were placed there by a beloved one rather than carried away from the neighbour's door by a strong gust of wind.

Attribution theory is concerned with how people make use of the information present in our physical and social environment to arrive at causal explanations for events. In this chapter, I will first provide a brief overview of attribution theory's historical roots by discussing the work of Fritz Heider, Harold Kelley and Bernard Weiner. Against this background, I will then focus more specifically on theories of responsibility attribution. I will draw a distinction between what I call *label theories* and *formal theories* of responsibility attribution. *Label theories* aim to capture the entirety of factors relevant to people's responsibility attributions but are less concerned with quantitative predictions of how the different factors influence attributions. Within the family of label theories, I will discuss normative models that prescribe how an ideal observer should make responsibility attributions and descriptive models that are more concerned with describing the processes by which people actually arrive at their attributions. *Formal theories*, in contrast, tend to focus on a subset of the relevant factors but make precise predictions about how this subset affects attributions of responsibility. Different formal theories

have employed different formal tools to model responsibility attributions, such as covariation or counterfactual dependence. In summarising the strengths and weaknesses of label theories and formal theories, I will argue for the importance of developing formal approaches but also note that previous attempts have been unsuccessful in providing a principled account of the relationship between causality, counterfactuals and responsibility. Such an account which conceptualises attributions of responsibility in terms of counterfactuals defined over a causal representation of the situation will be developed in Chapter 3 and tested empirically in subsequent chapters.

2.1 A Brief History of Attribution Theory

Attribution theory tries to explain how we come to understand why other people (and indeed we ourselves) behave the way they do (see Kelley & Michela, 1980, for a review of early work in attribution theory). Hence, it is a theory of common sense and has as its explanandum the causal attributions of naïve everyday people. Since it aims to explain the common-sense judgments of the folk, many people have the feeling that they already know attribution theory (Kelley, 1973). However, people also believe many things that are contradictory (e.g. 'same and same go together' versus 'opposites attract') or indeed simply false. Attribution theory tries to bring theoretical systematicity and clarity into the muddied waters of folk psychology. In this section, I will discuss the theoretical beginnings of attribution theory by focussing on the works of three very influential figures in the field: Fritz Heider, Harold Kelley and Bernard Weiner.

2.1.1 Fritz Heider – The father of attribution theory

The seminal work of Fritz Heider (1958) is generally regarded as the starting point for the scientific research into attribution. In his earliest work, Heider was concerned with perception as the primary way in which people come to an understanding of the physical and social world. He was intrigued by constancy phenomena in perception. The colour of my desk appears more or less the same to me quite irrespective of the time of the day – whether in the sunshine of early morning or under the light of my desk lamp at night. Similarly, my perception of the table's shape and size does not change as a function of where I am standing in the room. Despite the fact that the perceptual qualities of the table vary substantially – its size and shape as a function of where I am standing and its colour as a function of the time of day – it looks like the same old table to me. How can it be that such a high degree of variance in the proximal stimulus (the retinal image of the table) can give rise to a constant distal image of the object (the table out there in my room)?

While vision scientists have found out a great deal about the constancy of colour and shapes in recent years (see for example, Walsh & Kulikowski, 1998), Heider was

not only interested in why tables look the way they do but also in the question of why other people look the way they do. More precisely, he was interested in the question of how it can be that we look through the surface level of observable behaviour to the hidden depths of underlying dispositions, desires and intentions. Just like the multitude of different percepts on my retina give rise to a constant image of the table, so can a multitude of behaviours displayed by my housemate reveal a single underlying intention. If, for example, I hear that my housemate tries to open a door on the top floor, then a door on the second floor, I can be fairly confident that he or she has the intention to use the bathroom. I can then predict that he or she is likely to try and use the bathroom on the ground floor (and I could even make an educated guess of what is likely to happen if that door happened to be closed as well ...).

One of the main themes of Heider's work is that people are not content with the perceptual information in itself. As Cicero says, we strive for an understanding of the underlying causes that give rise to what we see – whether it is people, tables or small geometrical shapes frantically moving around a big square. The latter example was used to investigate our irresistible tendency to attribute mental states to others in order to explain their behaviour. The ease with which we generate mentalistic explanations (that is, explanations of behaviour in terms of mental states, such as beliefs, desires and intentions) was vividly demonstrated in Heider and Simmel's (1944) study (see Figure 2.1).

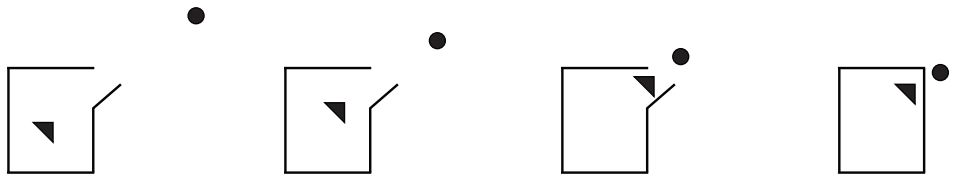


Figure 2.1: Figure adapted from Heider and Simmel (1944).

Participants viewed a short animation that depicted movements of geometrical shapes and were later asked to describe what they had seen. The animation showed a rectangle with a moveable part and some additional geometrical shapes. If one were to describe what happened in the animation solely in terms of the physical movements of the objects, one would (i) need a lot of words and (ii) render what happened mostly unintelligible. While one or two participants did in fact describe the clip just in terms of physical motion, the vast majority of participants in the experiment adopted what Dennett (1989) calls the *intentional stance*: they explained their observation of the complex system of moving geometrical shapes in terms of mental states. For example, to describe the clip in Figure 2.1, a prototypical response could be: "The circle *wants* to go in the *house* and *tries* to go through the *door*. However, the triangle does not *want* the circle to get

in. Maybe they don't *like* each other. The triangle *prevents* the circle from going in the *house* by *shutting the door* in front of its *nose*."

A multitude of mentalistic concepts and anthropomorphisms were used to describe the scene. Not only is such a description shorter and more intelligible than a mere recapitulation of the physical movements, but it also extracts important invariances in terms of the agent's motives in the interaction, that enable prediction and explanation of many other possible patterns of behaviour (or motion). For example, if it is true that the two shapes don't like each other, we can predict how they are likely going to behave in other circumstances such as when a diamond shape appears who happens to be good friends with the circle.

In Chapter 4 of *The naive analysis of action*, Heider (1958) lays the foundations of what is nowadays often referred to as the *man-as-scientist* metaphor. The core idea is that people, just like scientists, carefully observe and manipulate the world in order to reveal the underlying dispositional properties of their physical and social environment. Thus, in the same way in which a scientist conducts experiments and draws inferences about the invariant structural properties of her subject matter of interest, so do people act in their world so as to find out about the invariant properties of the world and other people around them (cf. Gopnik & Wellman, in press; Wellman & Gelman, 1992). The main reason for doing so – apart from sheer curiosity – is to develop a deep understanding of the world which allows one to *predict* what is likely to happen in the future, to *explain* why certain states of affairs came about and to *manipulate* the world in order to achieve the effects one desires (Hitchcock, 2012; Woodward, 2003).

Heider takes Kurt Lewin's (1936) famous equation $B = f(P, E)$, which states that behaviour B is a function of the person P and the environment E as a starting point for his theory of attribution. He then proceeds with a careful analysis of what factors influence whether the causes of a person's behaviour (or the outcome of this behaviour) are attributed to the environment or to the person. For example, if I observe that someone fails to solve a Sudoku puzzle on my tube ride, I can attribute this failure to causes that lie within the person, such as the person's lack of ability or effort. Alternatively, I can attribute the failure to causes that lie outside the person, such as the difficulty of the task or the noisiness of the tube. Indeed, almost any observable behaviour is a function of *internal* and *external* causes. The majority of the research in early attribution theory has been concerned with identifying the conditions that lead to internal or external attributions (Kelley & Michela, 1980).

But how do people identify what factors are causally responsible for the observed behaviour of another person given that any behaviour is always influenced by a multitude of causes (both internal and external)? As alluded to above, Heider's answer to this problem is that people proceed very much like scientists: he claims that the systematicity with which people arrive at their attributions is essentially an intuitive implementation of John Stuart Mill's (1898) method of experimental inquiry. Accord-

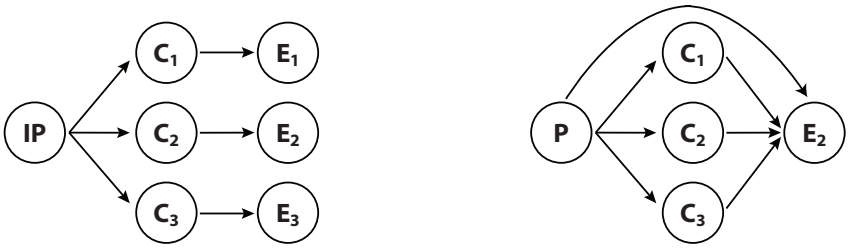
ingly, attributions proceed through covariational analysis: "We shall start with the data pattern fundamental in the determination of attribution, namely: that condition will be held responsible for an effect which is present when the effect is present and which is absent when the effect is absent."(Heider, 1958, p. 152)

With this in mind, let us now turn to what Heider had to say about the core theme of this thesis – how people attribute responsibility to each other. Heider was arguably first to formulate a theoretical model of responsibility attribution. In his model, Heider proposes that there are several levels of responsibility which differ with respect to how strongly a person and the outcome of the person's behaviour are linked. The stronger the linkage between outcome and person, the more responsible the person is for the outcome. Generally, Heider's theory of responsibility attribution is closely connected to his theory of action attribution: different conditions for action are predicted to result in different attributions of responsibility.

On the first level, the level of *association*, "the person is held responsible for each effect that is in any way connected with him or that seems in any way to belong to him" (Heider, 1958, p. 113). On this very general level, a football fan could be held responsible for his favourite team's win – it is sufficient that the fan is closely associated with the football players who are themselves more directly responsible for the outcome. On the second level, the level of *causality*, a person is seen responsible for an outcome for which her action was instrumental, whereby Heider analyses causation in terms of necessity – a person is responsible if the person's action was necessary for the outcome to occur.[1] Crucially, on this level, the person neither had foresight of the consequences of her action nor did she intend to bring about the outcome. Hence, a person is judged based on the outcomes she brought about rather than on her internal motives. Piaget (1932) refers to this level as *objective responsibility*. For example, if I open the cupboard in the kitchen and a plate falls out crashing on the kitchen floor, I would be causally (or objectively, in Piaget's terms) responsible. On the third level, the level of *foreseeability*, a person is held responsible for any effects that he caused and could have foreseen but did not intend. A doctor who administers a drug to a patient who subsequently has an allergic reaction to the drug would be held responsible on this level *if* the doctor could (or should) have foreseen this side-effect. On the fourth level, the level of *intention*, a person is held responsible only for those effects that she brought about intentionally. Piaget refers to this level as *subjective responsibility*.

Heider argues that intentions are the core feature that distinguishes personal from impersonal causation (see Figure 2.2). Impersonal causation is marked by a high dependence on environmental circumstances that determine to a large degree the nature of the effect. Hence, depending on the circumstances (C_{1-3}), different effects (E_{1-3}) will result (see Figure 2.2a). Personal causation (Figure 2.2b), in contrast, is marked

[1]Note that this definition of causality in terms of necessity conflicts with the earlier covariational account of causality.



(a) Impersonal causation **(b)** Personal causation

Figure 2.2: The difference between personal (P) and impersonal causation (IP). C_{1-3} are different environmental conditions, E is the effect. The direct arrow from P to E_2 indicates that E_2 is P's intended goal.

by what Piaget (1932) termed *equifinality*. The person intends to bring about a certain effect and will make sure that the same effect prevails independent of external circumstances. Hence, the relationship between the cause and the effect is much more robust in personal as opposed to impersonal causation.

Whereas in the case of impersonal causation, small perturbations in the situational conditions lead to different outcomes, the same outcome is brought about in the case of personal causation largely independent of differences in the situational conditions. The difference between personal and impersonal causation can be illustrated via the example of placing versus throwing a die. In the case in which I am placing the die on the table, I am fully responsible for the number that shows up on the die – this is an example of personal causation. Independent of external conditions, such as a wobbly table, I will bring about the intended outcome in a reliable fashion. However, if in contrast, I were to roll the die, I would be less responsible for the outcome. Here the outcome is strongly dependent upon situational factors and I lack the ability to bring about the same outcome in a reliable fashion. Although it is of course in principle possible that the same sequence of events comes about as a result of impersonal or personal causation (see the sequences $IP \rightarrow C_2 \rightarrow E_2$ and $P \rightarrow C_2 \rightarrow E_2$). This example shows that impersonal causality is not equated with non-personal or physical causality. According to Heider, a person is an impersonal cause in situations in which an intention to bring about the outcome is lacking.

Finally, on the fifth level, the level of *justification*, responsibility is not only a function of the person's intentions but also of the situational circumstances. Generally, a person's responsibility is a function of the relative contribution of (intra-)personal and external factors in bringing about the outcome. The more external factors were perceived as having contributed to the outcome, the less the person will be held responsible. For example, if Tom was coerced to assist in a bank robbery while his wife was kept as hostage, Tom's responsibility would be reduced despite the fact that he had formed the intention to assist in the robbery. In Heider's (1958) words, a person is not responsible if

"anybody would have felt and acted as he did under the circumstances ... responsibility for the act is at least shared by the environment" (p. 114).[2]

In sum, we see that Heider's conception of responsibility builds on an impersonal notion of causality and then successively incorporates mental states and aspects of the situation to arrive at a full-fledged model of responsibility.

2.1.2 Harold Kelley and Bernard Weiner – Two prolific sons

Both Harold Kelley and Bernard Weiner were heavily influenced by the work of Fritz Heider. Kelley took Heider's *man-as-scientist* metaphor and developed it into a formal model of causal attribution based on the covariation principle. Weiner focused on Heider's analysis of action and concentrated mostly on the personal and environmental factors that underlie causal attributions in achievement contexts.

2.1.2.1 Kelley's ANOVA model of causal attribution

Kelley (1967, 1973) took Heider's *man-as-scientist* metaphor as a theoretical starting point and proposed that in finding out about the underlying causes of another person's behaviour, people perform the naïve equivalent of an analysis of variance (ANOVA). Accordingly, "an effect is attributed to the one of its possible causes with which, over time, it covaries." (Kelley, 1973, p. 108)[3] Kelley distinguishes two fundamental principles of causal attribution: (i) the *covariation principle* and the (ii) *configuration principle*. The covariation principle only applies in situations in which repeated observations are available. However, Kelley acknowledges that people regularly make causal attributions based on single observations only. The configuration principle describes the conditions that people rely on when making attributions in the absence of repeated observations.

Let us first consider the covariation principle. According to this principle, people partition the events in their environment into causes (the independent variables in the ANOVA) and effects (the dependent variables).[4] Kelley (1967, 1973) proposes a model according to which information about *consensus*, *distinctiveness* and *consistency* are key dimensions which people use to arrive at their causal attributions. Depending on the information that people have about these different dimensions, they are predicted to either see the causal locus of the effect in the *stimulus*, the *person* or a combination of the *circumstances and the person*. How these different dimensions explain people's

[2] We will later see Heider's intuition formalised in a model that predicts that people's attributions of responsibility are influenced by the subjective degree of belief that someone else would have acted in the same way in the given situation (Fincham & Jaspars, 1983).

[3] Kelley's ANOVA model of causal attribution is one of the best historical examples of what Gigerenzer (1991) calls the tools-to-theories heuristic (see also Weiner, 1991). This heuristic describes scientists' tendency to use the tools that form part or their everyday scientific life as metaphors to inform their theories.

[4] The theoretically interesting question of how people achieve this partition is left unexplained. As we will see below, the inability of covariational accounts to distinguish between causes and effects is a major criticism of such frameworks.

attributions is best illustrated via an example. Consider the following pieces of information (cf. McArthur, 1972): Tobi is amazed by a concert of the band Gravenhurst. His housemate Renny and quite a few of the other people in the audience are not too impressed. Tobi often comes back from other concerts being somewhat disappointed. However, each time Tobi got back from previous Gravenhurst concerts, his eyes were still shining from excitement.

What was the reason for Tobi's being amazed by the Gravenhurst concert? Is it something about Tobi? Is there something special about Gravenhurst? Were the particular circumstances responsible for Tobi's amazement? Or does some combination of these factors explain why Tobi was so excited about the concert? According to Kelley's (1973) account, people should say that the the combination of the person (Tobi) plus the stimulus (Gravenhurst) are responsible for the observed effect. Let's see why.

Information about people, or *consensus* information, provides evidence about whether there is something special about the person. Does the effect (amazement) only occur in the person of interest or is it shared by other people as well (e.g. Renny)? In our case, consensus was relatively low as the amazement was not shared by all people. *Distinctiveness* information describes whether the effect is distinct with respect to the particular stimulus or is shown across a broad range of comparative stimuli. Tobi particularly liked the Gravenhurst concert but he does not like all concerts. Hence, distinctiveness was high. This tells us that there is something special about Gravenhurst concerts (at least for Tobi). Finally, *consistency* information relates to whether there is something special about the particular circumstances. Does Tobi always show the same behaviour when going to Gravenhurst concerts or only sometimes? We have learned that he enjoyed it several times before and hence we can be confident that consistency is high.

The idea that causal attributions are determined via a systematic consideration of the consensus, distinctiveness and consistency dimensions has not remained uncontested. Hilton and Slugoski (1986), for example, have argued that people's causal attributions are guided by considerations about how people normally behave in particular situations rather than following the predictions of the covariation analysis. A potentially more fundamental problem with the covariation approach which was already noted by Heider and Simmel (1944) is that people often make attributions without having the information available on which the ANOVA model relies. "Generally, we don't postpone attribution until we tally a series of joint condition-effect changes. We judge the *book* to be good, the *food* to be excellent, the *child* to be nice, the *scene* to be beautiful, because we experienced pleasure upon the specific and single contact with it." (Heider, 1958, p. 155)

This is where Kelley's second fundamental principle of attribution, the *configuration principle*, comes into play. When people do not possess sufficient covariational information, they are assumed to rely on beliefs about the exact relationship between the perceived causes and the effect (Kelley, 1983). What specific configuration of factors has

caused the effect? How do the different potential causes combine in order to bring about the effect? Kelley (1972) calls these configurations of factors *causal schemas*. "A schema is derived from experience in observing cause and effect relationships, from experiments in which deliberate control has been exercised over causal factors, and from implicit and explicit teachings about the causal structure of the world." (Kelley, 1973, p. 115) Once acquired, a causal schema can be used to draw inferences from single observations based on causal configurations. In Kelley's (1972) general framework, a schema can be thought of as a collection of previous observations that enters the ANOVA together with the current observation.[5]

Two examples for schemas are the *multiple sufficient causes* (MSC) schema and the *multiple necessary causes* (MNC) schema. The MSC schema captures a configuration in which multiple causes are individually sufficient to bring about the effect. In the MNC schema, in contrast, the presence of each cause is necessary for the effect to occur. The two schemas have different implications for what inferences can be drawn from observing that the effect (and some of the causes) occurred. Consider the situation in which two men, Jack and Bill, play chess. Bill wins the game. What caused Bill's win? In terms of the MSC schema, we could say that there are two sufficient causes for Bill's win. Either Bill is strong or Jack is weak. If one or both of these conditions are met, then Bill wins. From observing that Bill won, we cannot be sure which of these two conditions is responsible for his win. However, if we additionally happened to know that Jack has lost most of his previous games, we would be less inclined to think that Bill is a strong player. Knowing that one of the sufficient causes was present (Jack is weak), explains away the effect and casts doubt on the presence of the other plausible cause (Bill is strong) is in doubt. Kelley calls this general pattern of inference the *discounting principle* (see Morris & Larrick, 1995, for a formal analysis of discounting). "The role of a given cause in producing a given effect is discounted if other plausible causes are also present."(Kelley, 1973, p. 113) The same example can also be used to illustrate another pattern of inference: the *augmentation principle*. The augmentation principle comes into play when one of the causes is assumed to be inhibitory. If, for example, we happened to know that Jack is a very strong player, we would infer that Bill must be even stronger. Finally, we could also cast the situation in terms of the MNC schema. Let's assume that Bill is a little boy and Jack a man in his forties. Because we are generally inclined to think that children are less skilled than adults (especially in chess), it appears that *both* an exceptional skill on behalf of Bill and a relatively poor ability of Jack are necessary to explain why Bill won the game.

Arguably, the three examples mentioned above might better fit into a *compensatory causal schema* which allows for the different causes to vary in strength in a quantitative

[5]The important question of how people decide what schema to apply in a given situation is again mostly left open.

manner (rather than just being absent or present).[6] However, the examples still serve to illustrate the general point: different causal schemas imply different causal attributions, whereby the schemas are subjective in the sense that they capture how the attributor thinks the causes combine. Given that we know what causal schema was evoked, we can predict the inferences a person will make for multiple possible patterns of evidence. In Chapter 3, we will see that a richer theory of causality will help us to be more precise about what schemas or models people use to represent a situation.

2.1.2.2 Weiner links responsibility, motivation and emotion

We have seen that Kelley was mostly concerned with analysing how people determine the causes of observed behaviour. Weiner, in contrast, was predominantly interested in the consequences of causal attributions. He developed a broad theoretical framework according to which causal attributions are the glue that combines disparate psychological phenomena such as responsibility, motivation and emotion.

Weiner develops his theory in the context of achievement tasks and closely follows Heider's analysis of people's concepts of *try* (effort) and *can* (ability). He begins with an identification of the most salient causes that people perceive in achievement contexts, namely, ability, effort, task difficulty and luck (Weiner et al., 1971). In contrast to Kelley's ANOVA model, Weiner suggests a somewhat different set of core dimensions that are supposed to underly causal attributions. He proposes that success or failure in achievement tasks can be understood in terms of the *locus*, *stability* and *controllability* of causal factors. The locus dimension describes whether the person or the task is the cause for failure/success. The stability dimension describes whether the causal source of the task outcome is invariant over time. Finally, the controllability dimension indicates to what extent the person exhibits control over the conditions for bringing about the outcome in a reliable fashion.[7]

In contrast to Kelley, Weiner does not focus so much on how people locate the causes of an observed effect but aims to explain the cognitive, emotional and behavioural consequences that result from causal attributions. Weiner proposes that causal ascriptions, which can be analysed in terms of the three outlined dimensions, are fundamental to understanding motivation. He demonstrates the importance of causal ascriptions in terms of their role for explaining how people change their expectations about future success after having failed or succeeded in a task. Generally, it was assumed that people increase

[6] Although see Kun and Weiner (1973) who have also explained causal inferences in achievement contexts in terms of the MSC and MNC schemas.

[7] Although Weiner distinguishes his model from Kelley's (1973) ANOVA model, the two theories are similar in important respects. Frieze and Weiner (1971) argue that the causes for success or failure are often sought in the difficulty of the task, the ability of the person or luck. In Kelley's terms, these would be attributions to the stimulus (i.e. the task), the person (i.e. the person's ability) or the combination of person and circumstance (i.e. luck, see Lipe, 1991). However, whereas Kelley's dimensions reveal information about the possible source of the cause but are themselves not causal, Weiner's dimensions describe *causal* properties of the relevant factors.

their expectancy of future success after having succeeded and lower their expectation after having failed.

However, this general pattern does not provide strong evidence for an attribution account since it can also be accounted for by a simple reinforcement model (Sutton & Barto, 1998). Interestingly, it was found that sometimes the opposite pattern occurs: a lowering of the expectation after success and an increase after failure. An attributional framework can shed light on when this effect is likely to occur. If, for example, the failure in a task is attributed to external and unstable factors and the person perceives the outcome to be in principle controllable, then a failure can be followed by an *increase* in the expectation of success. Consider the situation in which a skilled archer just misses the target because of a strong gust of wind. The archer might well increase his expectation of success on a subsequent attempt given this observation. First, the gust of wind is only a very temporary occurrence and will most likely not interfere on the next attempt. Second, his perception of the task's difficulty might have changed after the first shot. Given that he only just missed the target despite the gust of wind, he might now think that the task is even more controllable than he had originally thought and accordingly raise his expectations of a future success.

Weiner proposes that causal attributions not only have important consequences for motivation but also for people's emotional responses to different outcomes. Indeed, many theoretical frameworks of emotion include causal *appraisal* as a core component (Lazarus, 1966; LeDoux, 2000; Schachter & Singer, 1962). With *appraisal*, emotion theorists mean the appreciation of the sources of one's initial emotional response. For example, if a cyclist bumps me from the back as I am waiting in front of a red light, I might experience an initial emotion of anger towards the person. However, if I then turn around and see the other person apologetically pointing to his brakes that just broke, my emotion of anger will swiftly turn into pity. Not only did the person not intend to bump into me but also is she now in a tricky situation because cycling in London without brakes is not a pleasant experience.

Weiner analyses different emotions such as happiness, sadness, anger, pride and pity. While some emotions such as happiness and sadness turn out to mostly depend upon the valence of the outcome and are largely independent of how the outcome came about, other emotions such as anger, pride and pity are closely linked with causal attributions. We have seen in the example with the cyclist how anger can quickly become pity once one has appreciated the causal factors that led to the negative outcome. More generally, Weiner claims that once the configuration of causal factors is understood, it can be predicted what emotions are likely to be experienced. Cognition and emotion are intimately linked: "increasing cognitive involvement generates more differentiated emotional experience" (Weiner, 1985, p. 560).

Finally, Weiner (1995) has also analysed how the process of responsibility assignment features in the more general framework that links causal attributions with emotion

and motivation. Figure 2.3 gives an overview of the responsibility assignment process. It depicts the antecedents of responsibility assignments as well as the emotional and behavioural consequences. According to Weiner (1995) there is a sequence of conditions that have to be met in order for responsibility to be assigned to an event. The sequence describes the way in which a person is assumed to look for information in order to determine whether responsibility should be assigned.[8]

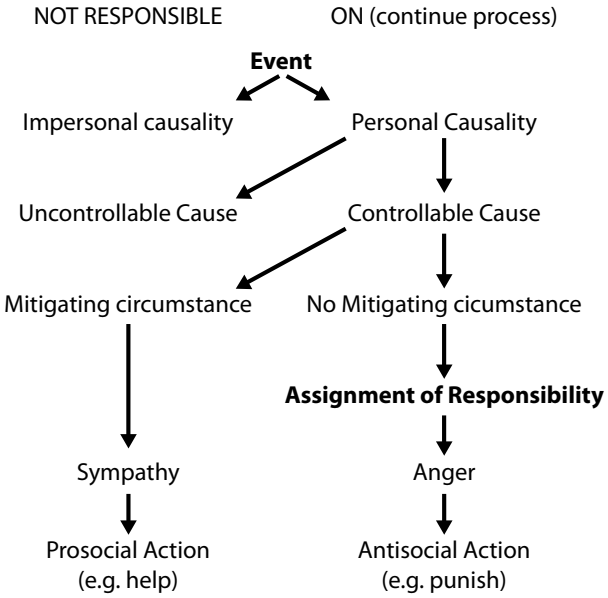


Figure 2.3: The responsibility assignment process with its emotional and behavioural consequences according to Weiner (1995).

Consider the following example: Ben is waiting in front of a red traffic light when something hits his back tire and almost knocks him off the bike. The first criterion that has to be met for an assignment of responsibility to be appropriate is that the causal event is personal. Here the notion of personal versus impersonal causality is different from Heider (1958) and merely distinguishes whether the event was caused by a person or not. If the cause is impersonal (e.g. a branch that fell from a tree and hit Ben's bike) then no responsibility is assigned. If the cause is identified as personal (e.g. another cyclist who crashed into Ben's back tire) then the responsibility assignment process continues. The second criterion which needs to be fulfilled is that the cause event was under the control of the person. If the event was uncontrollable (e.g. the brakes of the other cyclist just broke) then no responsibility is assigned. Finally, Ben is presumed to look for whether any mitigating circumstances exist that explain the occurrence of the causal event. For example, the other cyclist might offer an excuse and say that he is in a rush because his pregnant wife just called to tell him she has contractions. However,

[8]While the depicted process can be seen as an ideal path of information acquisition, Weiner acknowledges that the sequence is not necessarily invariant. That is, the order in which different conditions for responsibility assignment are assessed can vary between situations.

if no mitigating circumstances are apparent and the other cyclist does not offer any excuses, he is assigned responsibility for the event.

Overall, the responsibility assignment process is presumed to link causal attributions (the antecedents) with emotions and motivation (the consequents). If the conditions for responsibility are met, then a negative event elicits the emotion anger which triggers the motivation for antisocial behaviour. Ben is likely to become aggressive when having been hit by the other cyclist and could, for example, kick his bike. In contrast, if there are mitigating circumstances, Ben is assumed to feel sympathy towards him which in turn triggers a prosocial motivation. For example, if Ben found out that the other cyclist is in a rush because his pregnant wife is expecting, Ben might offer his help. Figure 2.4 depicts the chain that links the perception of an initial event with the resulting behaviour. The perception of an event elicits a causal search which results in a causal attribution. Depending on the causal attribution, responsibility is assigned or not. Whether responsibility is assigned or not then determines which emotion is felt. The emotion triggers a certain motivation which then culminates in the behaviour of the agent.

Figure 2.4: The causal chain that links the perception of an event with the resulting behaviour according to Weiner (1995).

As as can be seen in Figure 2.3, Weiner's theory of the responsibility assignment process neither features an assessment of intentionality nor foresight as preconditions for responsibility. However, despite the fact that they are not directly incorporated into his framework, Weiner acknowledges the importance of both intentions and foresight for assignments of responsibility. A person is held highly responsible if he "... wanted to perform the socially inappropriate behavior, engaged in the conduct with foresight and knowledge of its consequences, and may even have pursued a variety of means that were responsive to 'evasive' actions." (Weiner, 1995, p.13)[9]

Furthermore, Weiner draws a theoretical distinction between causal attribution and assignment of responsibility. Whereas the former is thought to be tightly linked to the nature of the event, the latter is more closely related to an assessment of the person. In Weiner's (1995) words: "The responsibility inference process is presumed initially to focus on causal understanding and then to shift to a consideration of the person." (p. 8-9). Hence, similar to other theorists (see Jones & Davis, 1965; Trope, 1986) responsibility assignments are based on an inference about the dispositional mental states of the agent from the perceived acts. However, Weiner sees the motivation for why people infer the mental states of others as not only being driven by their 'scientific'

[9]The mentioning of the multiple means to achieve the pursued goal is reminiscent of the notion of equifinality discussed above (see Figure 2.2).

curiosity. He advances a novel metaphor: *man-as-lawyer* (see Fincham & Jaspars, 1980; Hamilton, 1980).[10] Accordingly, people are not only interested in explaining why people act the way the do but are also concerned that justice prevails. Others are seen to deserve blame and punishment for their wrongdoings and praise and reward for honourable deeds.

2.1.3 Summary

In this section, I have provided a short summary of the origins of attribution theory from Fritz Heider to Harold Kelley and Bernard Weiner. Heider emphasised people's irresistible tendency to look for the causes of things. He also developed the first psychological model of responsibility attribution in which he highlighted the importance of intentionality and the ways in which behaviour is shaped by both internal and external causes. Kelley has further elaborated Heider's *man-as-scientist* metaphor and proposed that people use covariational and configurational principles to discern the causes of observed behaviour. Finally, Weiner has highlighted the importance of causal attributions for motivation in achievement contexts and for explaining what emotions people are likely to experience in different circumstances. He also developed a model of responsibility in which the notion of control features prominently. Having covered the major frameworks in early attribution theory, we will now turn to theoretical frameworks that have focused on attributions of responsibility in particular.

2.2 Label Theories of Responsibility Attribution

In this section, I will discuss a family of responsibility attribution theories which I call *label theories*. Label theories are mostly concerned with identifying the factors that people use (or should use) when making attributions of responsibility. However, they do not provide a precise formulation of exactly how the different factors are predicted to influence attributions. I will first discuss normative theories that specify how an ideal observer *should* attribute responsibility and then discuss descriptive theories which aim to explain how people *actually do it*. After having concluded this section by discussing the strengths and weaknesses of label theories, we will look at theories that aim to provide a more formal treatment of the attribution process.

2.2.1 Normative theories

In his book *The attribution of blame: Causality, responsibility and blameworthiness*, Kelly Shaver (1985) develops a comprehensive theory of blame attribution. His theory

[10] Additional metaphors that have been thrown in the mix include: *man-as-theologian*, *man-as-politician*, *man-as-economist* and *man-as-prosecutor* (see Tetlock, 2002).

builds on the work of social psychologists, such as Heider (1958) and Kelley (1973) but also pays close attention to the work of legal scholars (cf. Hart & Honoré, 1959/1985).

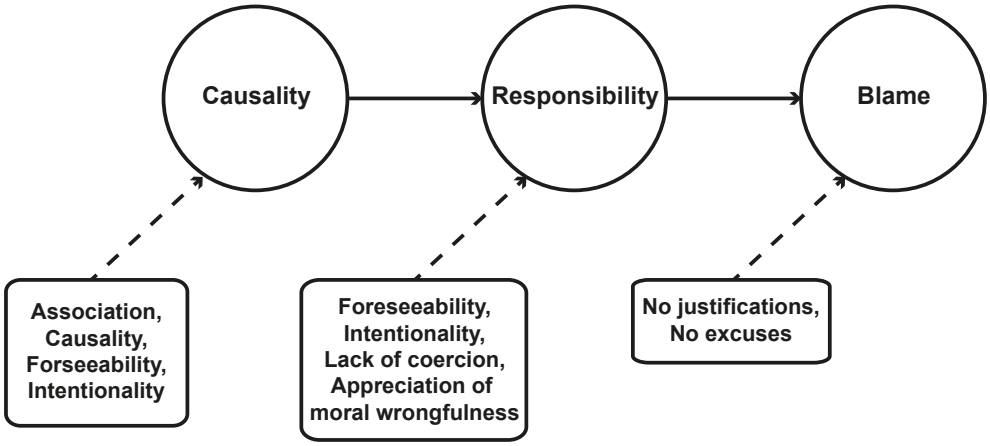


Figure 2.5: Shaver's (1985) normative theory of blame attribution.

Figure 2.5 shows a graphical representation of Shaver's (1985) theory. The theory is *normative* in the sense in that is describes how an ideal observer should arrive at an attribution of blame in a given context. Shaver distinguishes between three core concepts: *causality*, *responsibility* and *blame*. Several conditions have to be met for each of the corresponding attributions to be justified. Similar to Heider (1958), Shaver describes a sequential process in which each of the conditions for an earlier step have to be met to reach a later step in the process. Hence, the concepts stand in a relation of presupposition: there is no responsibility without causality and no blame without responsibility.

In his analysis of *causality*, Shaver follows Heider (1958) very closely. Causality increases with the extent to which the agent is linked with the outcome. The strength of the linkage of agent and outcome increases from situations in which agent and outcome are merely associated to situations in which the agent brings about the outcome intentionally. Shaver also includes causality and foreseeability as intermediate steps.

Given that the conditions for causality are met, the ideal attributor should proceed to deciding whether it is appropriate to attribute *responsibility*. For attributions of responsibility, the agent's foresight as well as whether the agent brought about the outcome intentionally play key roles. We see that that the concepts of foreseeability and intentionality appear twice in Shaver's model: first, as conditions for causal attributions and second, as conditions for attributions of responsibility. However, Shaver argues that the role these factors play are different depending on whether they are considered for attributions of causality or responsibility. While their influence for attributions of causality is expected to be relatively minor, attributions of responsibility are thought to be much more strongly dependent upon these mental states of the agent. Furthermore,

attributions of responsibility should be sensitive to mitigating circumstances. If, for example, the agent was coerced to bring about the outcome, attributions of responsibility should be reduced if not absent. Finally, Shaver also adds the criterion that the agent should be able to appreciate the moral wrongfulness of his action in order to be held responsible. To the extent that an agent is unable to appreciate the moral consequences of their action, responsibility should be reduced.

When all the conditions for an attribution of responsibility are met, the ideal attributor should proceed to assess whether an attribution of *blame* is appropriate or not. Shaver distinguishes blame from responsibility through two additional requirements: a person is only blameworthy if he is unable to offer a justification or an excuse for his behaviour. Thus, if the person himself is able to justify his behaviour or the attributor learns about additional factors that serve as valid excuses, the person should not blamed even though he is still responsible for the outcome. In order to excuse oneself, a person can, for example, deny the foreseeability of the negative outcome (see Darley & Zanna, 1982; Markman & Tetlock, 2000)[11], highlight the presence of coercive factors (see Hamilton, 1980) or downplay the causal influence his action had on the outcome (see Kerr & Kaufman-Gilliland, 1997).

While Shaver's theory is still unparalleled in its attempt to provide a comprehensive psychological treatment of the process of blame attribution, there are some obvious problems with his account (cf. also Lagnado & Channon, 2008). First of all, his conception of causality is problematic. The inclusion of 'causality' as a factor for causal attribution is circular. Furthermore, it is unclear why mental state variables such as intentions and foresight *should* influence causal attributions. While in Heider's (1958) theory they served as different levels for the attribution of responsibility and to distinguish personal from impersonal causation, it is unclear what role they play in Shaver's theory. We will see in the next chapter that theories of causality have advanced considerably in the meantime (Pearl, 2000; Woodward, 2003). I will argue that these richer theories of causality provide a stronger theoretical background for theories of responsibility and blame attribution.

Shultz and Schleifer (1983) have also proposed a normative model of the attribution process that is quite similar to Shaver's (1985) model discussed above. Figure 2.6 shows an overview of their model. Due to its similarity with Shaver's theory, I will just highlight a few facts without going into the exact details. Shultz and Schleifer (1983) conceptualise causality as a combination of production with logical dependence. Hence, a person is seen as having caused the outcome if they brought it about and/or if their contribution was sufficient/necessary for the outcome to occur. In contrast to

[11] Whether this will count as an excuse will depend to some degree on whether the outcome *should* have been foreseen. For example, whether a doctor should be blamed who did not foresee the negative side effects of some medication, depends on whether being aware of the negative side effects is something to be expected from a person with that profession. This is reminiscent of the criterion of a *reasonable man* which is often employed as a test for legal liability (see Green, 1967).

Shaver (1985), Shultz and Schleifer (1983) reserve mental state variables to attributions of responsibility and do not analyse causality in terms of mental states as well. Moral responsibility is contingent on whether the agent foresaw and intended the outcome. Furthermore, it matters whether he was free to act and did so in a negligent or reckless manner. Finally, a person should only be held morally responsible if there were no other alternative causes that intervened to bring about the outcome in a way that broke the causal process initiated by the agent's action (Fincham & Shultz, 1981).

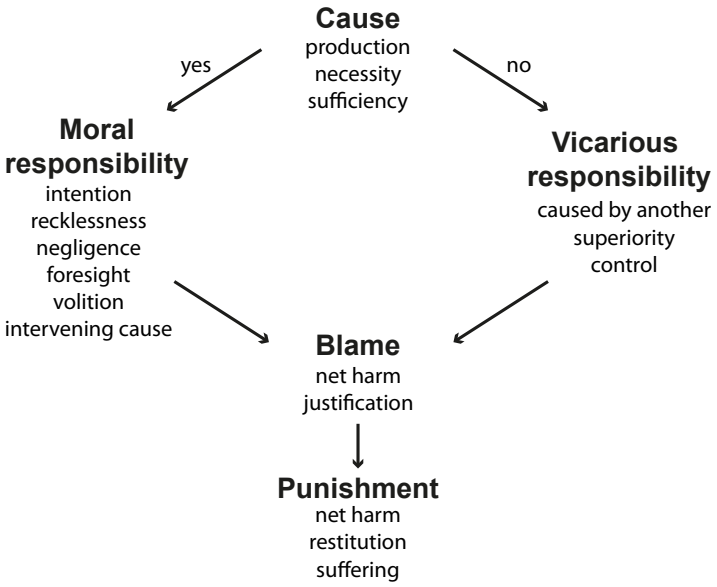


Figure 2.6: The relationship between causation, responsibility, blame and punishment according to Shultz and Schleifer (1983).

Shultz and Schleifer (1983) also explicitly acknowledge the possibility of vicarious responsibility in situations in which the negative outcome was caused by another person (see Shultz, Jaggi, & Schleifer, 1987). Conditions for vicarious responsibility are that the person to be held responsible is in a superior position (such as a manager in a company or a parent in a family) and exerted (potential) control over the subordinate's behaviour.

Blame is distinguished from responsibility in a similar vein as in Shaver's model by considering whether the person can offer justifications for their behaviour. Furthermore, they include the additional condition that blame is only appropriate when the behaviour of the person of interest caused more harm than benefit to the victim. Finally, if the person is considered blameworthy, punishment can be inflicted upon him. Shultz and Schleifer (1983) argue that punishment is related to the amount of suffering experienced by the victim and the extent to which restitution is appropriate for compensating the victim.

Both Shaver (1985) and Shultz and Schleifer (1983) take great care to distinguish

the concepts of causality, responsibility and blame from each other. Both argue that it is theoretically important for researchers to be clear about what concept they are investigating (Shaver & Drown, 1986). The problem of distinguishing between the concepts is aggravated by the flexibility with which people use them in everyday talk. We can, for example, blame the rain for a failed picnic attempt (an example that comes readily to mind in London). Of course, the rain meets hardly any of the conditions that normative theories consider necessary for an attribution of blame. Hence, no matter what theoretical demarcations one will draw as a researcher, it is unlikely to be possible to establish a direct mapping to people's linguistic usage of the different terms. However, this does not mean that it is in vain for researchers to make some effort in distinguishing the different concepts theoretically.

2.2.2 Descriptive theories

It will come as no surprise to the reader that people's responsibility attributions do not always follow the prescriptions of the normative theories discussed in the previous section. Work in social psychology has identified a host of biases that show how people's actual attributional verdicts often diverge from what would be expected from a normative perspective. People have a tendency to attribute behaviour to a person's disposition and neglect the presence of situational constraints (*fundamental attribution bias*, cf. Gilbert & Malone, 1995; Jones & Davis, 1965; Jones & Harris, 1967; Ross, 1977), to explain one's own behaviour in terms of situational and other's behaviour in terms of personal factors (*actor-observer bias*, cf. Jones & Nisbett, 1971) and to take credit for positive outcomes but deny blame for negative ones (*self-serving bias*, cf. Bradley, 1978; Miller & Ross, 1975). What is worse, we tend to look at the speck of sawdust in someone else's eye but pay no attention to the plank in our own eye. We think of others as biased attributors but are mostly unaware of our own biases (Kruger & Gilovich, 1999; Pronin, Gilovich, & Ross, 2004).

Whether a certain attribution qualifies as biased depends of course on the adopted normative framework. There is considerable disagreement about whether or not particular patterns of attribution should be viewed as biased (Forsyth & Kelley, 1994; Hamilton, 1980; Malle, 1999, 2006; Mezulis, Abramson, Hyde, & Hankin, 2004; Miller & Ross, 1975; Rantilla, 2000; Ross & Sicoly, 1979; Schlenker & Miller, 1977). For example, consider a situation in which several men harass a foreigner. From the perspective of Kelley's (1973) ANOVA model, it would count as an error to attribute one of the men's aggressive behaviour to that man's disposition. The fact that other men behaved in the same way should favour an attribution to the 'stimulus' (i.e. the foreigner) rather than the person (especially if we happen to know that the man hasn't been aggressive before). However, when an alternative normative standard for evaluating a person's behaviour is adopted, an inference to the person's disposition might

indeed be valid.

Hamilton (1980) has argued that people are not only *intuitive psychologists* but also *intuitive lawyers*. Intuitive psychologists and intuitive lawyers follow different purposes in their attributions. The *psychologist* aims to find an explanation for a surprising effect and searches for the underlying cause (Kelley, 1967, 1973). The *lawyer*, in contrast, is concerned with whether it is appropriate to socially sanction an action (Jones & Davis, 1965). Hamilton suggests that *psychologists* and *lawyers* not only follow different purposes but also employ different decision rules: whereas *psychologists* use the covariation principle, *lawyers* use a 'could have done otherwise' test (Hamilton, 1980, p. 768). Accordingly, whether or not a person is held responsible for the outcome depends on whether the attributor thinks that the person could have acted otherwise and that acting otherwise would have resulted in a better outcome. Thus, while from an *intuitive psychologist's* perspective an attribution to the person might be biased, from the alternative viewpoint of the *intuitive lawyer* who thinks that the man should (and could) have acted otherwise, a personal attribution is appropriate.

2.2.2.1 Alicke's (2000) culpable control theory of blame

Alicke (2000) has developed a descriptive theory of blame attribution that describes different processes through which our attributions are influenced by motivated reasoning processes. Recall from our discussion of the normative frameworks of responsibility attribution above that several researchers have argued that the relationship between causality, responsibility and blame is one of presupposition (see, e.g Darley & Shultz, 1990; Shaver, 1985; Shultz, Schleifer, & Altman, 1981). Thus, the ideal observer should first make a causal assessment of the situation, then decide whether an attribution of responsibility is appropriate based on inferences about the persons' mental states and finally attribute blame if the person cannot offer any convincing justifications or excuses. Alicke (2000) argues that this ideal-observer model (while maybe an appropriate normative standard) does not reflect how ordinary people make attributions. Consider the following two versions of a scenario (adapted from Alicke, 1992):

> John was driving over the speed limit in order to get home in time to *hide an anniversary present for his parents / hide a vial of cocaine* that he had left out in the open before they could see it. As John came to an intersection, he failed to see a stop sign that was covered by a large tree branch. As a result, John hit a car that was coming from the other direction. John hit the driver on the driver's side, causing him multiple lacerations, a broken collar bone, and a fractured arm. John was uninjured in the accident.

How responsible do you think John was for the accident? Does it matter for how responsible he is whether his motive was good (hiding a present) or bad (hiding cocaine)?

Participants in Alicke's (1992) experiment thought so. John was seen as significantly more responsible for the accident when his motive was bad rather than good. The valence of John's motive also affected to what extent participants rated John to be the cause of the accident.

This set of findings is problematic for the normative models discussed above. According to these models, what John's motive was should not really make a difference as to how much he was seen as having *caused* the accident. Causal attributions are supposed to be made prior to moral evaluations. However, Alicke's (1992) results suggest that people's moral evaluations of John influenced how they conceived of his causal status in the scenario despite the fact that his relevant mental states, such as his intention to get home as quickly as possible and the foreseeability of an accident as a result of speeding, were the same in both stories and independent of his motive.

One of Alicke's (2000) key points is that the order in which people evaluate different factors before making their attributions is in stark contrast to the one prescribed by normative theories. Attributions of responsibility are directly triggered by the perception of a positive or negative event from which we reason diagnostically to try and find the causes. The perception of a significant event, especially when it was negative, elicits spontaneous emotional and moral evaluations. These evaluations then in turn lead to what Alicke refers to as *blame-validation processing*. Accordingly, on experiencing a negative event, we spontaneously feel that someone is to be blamed, which leads us to perceive the circumstances in which the negative event occurred in such a way that our feeling to blame someone is validated.

Knowing about John's motive affects whether or not the blame-validation-process is initiated. When we think of John as a good person, we don't want to blame him and look for mitigating circumstances that exculpate him. However, when we think of John as a bad person, we think that he *should* be blamed for the accident. Alicke points out several ways in which the blame-validation process might operate: we engage in confirmatory information search and look for evidence that supports our initial hypothesis of the agent's blameworthiness (Nickerson, 1998). Alternatively, we may change our evidential standards and stipulate that John *should* really have foreseen what was coming, ignoring the fact that the branch of the tree actually made it virtually impossible for him to see the other car. Finally, we can change our perceptions of the extent to which John had control over the outcome.

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. Alicke analyses how much control a person has over the outcome in terms of three structural linkages between mental, behavioural and consequence elements (see Figure 2.7). First, the mind-to-behaviour link captures the degree to which the actor's behaviour was perceived to have been under volitional

control. *Volitional behaviour control* is predicted to increase with the belief that the person acted purposefully and knowingly. It is diminished by situational constraints (e.g. coercion) and capacity constraints (e.g. the inability to correctly evaluate a situation or to act appropriately).

Second, the behaviour-to-consequence link characterises 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?). Thus, causal control is predicted to be diminished when a multitude of causes were present, when the cause event of interest was only remotely connected with the outcome and when the situation was such that the same outcome would have prevailed even if the cause had been absent.

Third, the mind-to-consequence 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. A doctor giving a wrong treatment to a patient who dies as a result, is an example of a situation with diminished volitional outcome control, assuming that the negative consequences of the treatment were in fact not foreseeable and that the doctor did not want the patient to die.

The blame validation process operates on each of the three structural linkages by systematically counteracting the factors that should normally result in a diminution of the actor's control over the outcome.

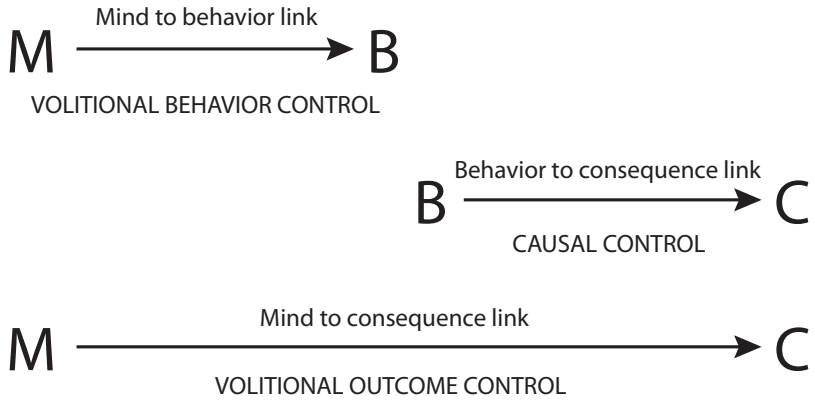


Figure 2.7: Structural linkages between mental element (M), behavioural element (B) and consequence element (C) according to the culpable control theory of blame (Alicke, 2000).

2.2.2.2 Experimental philosophy

More recently, work in the blossoming field of experimental philosophy (see, e.g., Knobe & Nichols, 2008; Sinnott-Armstrong, 2008) has issued an arguably even deeper threat

to normative models that conceptually separate attributions of causality from moral evaluations. Knobe (2006, 2009, 2010) and others (Driver, 2008; Hitchcock & Knobe, 2009; Knobe & Fraser, 2008) have argued that questions about non-moral concepts such as whether a person has caused (Hitchcock & Knobe, 2009) or intended (Knobe, 2003a) a particular outcome are not only influenced but *intrinsically linked* with moral evaluations. According to the strong view of this hypothesis, there is no such concept of causation (or intentionality) that is independent of moral considerations. For example, Knobe and Fraser (2008) had participants read the following scenario:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mails them reminders that only administrators are allowed to take the pens. On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionists desk. Both take pens. Later that day, the receptionist needs to take an important message ... but she has a problem. There are no pens left on her desk.

Participants were then asked to indicate whether they agreed or disagreed with each of the following two sentences:

1. Professor Smith caused the problem.

2. The administrative assistant caused the problem.

The results showed that participants tended to agree with the first statement but disagree with the second. These results are again problematic from a certain normative perspective. The actions of Professor Smith and the assistant are exactly the same – they each take one of the pens. The situation was described in a way that highlights that *both* the action of Professor Smith and the assistant were necessary to bring about the negative effect. The only factor that was different between the two actors concerns whether they acted with or against a norm. However, why should this matter for a judgment of causation? Why do people tend to select Professor Smith as the cause and not the assistant?

Hitchcock and Knobe (2009) argue that people's judgments of actual causation are intertwined with normative considerations (see also Halpern & Hitchcock, forthcoming). Judgments of actual causation refer to the selection of events as causes that brought about an effect in a particular situation. Inspired by Hart and Honoré (1959/1985) and Kahneman and Miller (1986), Hitchcock and Knobe (2009) argue that which events people select as causes of an effect in a particular situation is affected by counterfactual reasoning. In order to undo the outcome in the pen scenario above, it is sufficient to

either undo Professor Smith's action *or* the assistant's action. However, the two different counterfactual worlds differ in their availability: the world in which Professor Smith had not taken the pen comes to mind more easily because he is not supposed to take a pen in the first place (see Roese, 1997, for an overview of factors that have been identified to influence the generation of counterfactuals).

Norms influence which counterfactuals people are likely to consider which in turn influences which events they pick as the actual causes of an outcome in a particular situation. Hitchcock and Knobe (2009) identify and test a range of norms that are predicted to influence which counterfactuals people are likely to consider: moral norms (what should be done), statistical norms (what people tend to do) and norms of proper functioning (how things are supposed to work).[12]

Alicke, Rose, and Bloom (2011) offered a reinterpretation of Hitchcock and Knobe's (2009) findings. Rather than seeing the results as support for the view that norm violations per se guide causal attributions, Alicke et al. (2011) argue that actions against a norm serve as evidence about the moral character of the person which in turn influences causal attributions through the mechanism of blame validation as described above (see also Driver, 2008, for a discussion of the differences between Alicke's and Knobe's position). To test their alternative account, they created cases in which an actor who either acts against or according to a norm brings about a positive or a negative outcome. Whereas Hitchcock and Knobe's (2009) account predicts that behaviours which violate a norm should generally be seen as more causal, Alicke et al.'s (2011) model predicts that causal attributions should be moderated by the outcome (which influences blameworthiness). Thus, a person's perceived blameworthiness should have a direct effect on people's causal attributions which are not necessarily mediated by whether or not the person's actions were normative (although acting against a norm is often an indication for a person's negative character). In line with their predictions, Alicke et al. (2011) found that an actor can be seen as *less* causal for a negative outcome when having acted *against* a norm (compared to when he acted in line with the norm). Acting against a *bad* norm will make an actor look positive and praiseworthy such that his causal influence in bringing about a negative outcome will be downplayed.

To conclude, while there is evidence that norms influence attributions of causality, the simple view that non-normative actions always attract increased causal attributions appears mistaken. A person acting in line with a norm can be seen as more causal than a person acting against a norm (see Alicke et al., 2011; Sytsma et al., 2012). The

[12]Sytsma, Livengood, and Rose (2012) have recently qualified Knobe and Fraser's (2008) findings. They distinguish between *population-level norms* (what is the norm for a particular group of people, cf. Kelley's, 1973, consensus dimension) and *agent-level norms* (what a particular person tends to do, cf. Kelley's, 1973, distinctiveness dimension). In different variations of the pen scenario, they found that participants' causal attributions were in fact *not* influenced by population norms. Instead, causal attributions were higher for agents who acted according to their individual norm: that is, the assistant who almost always takes the pen (and who is allowed to do so) was seen as more causal for the outcome than the professor who almost never takes a pen (and who is not supposed to do so).

question of how strong the influence of normative considerations on causal attributions is remains an interesting area of research. Danks, Rose, and Machery (in press) have argued against the strong view which proposes that all of causal cognition (incl. causal perception, causal learning, causal reasoning) is inherently evaluative (cf. Knobe, 2009, 2010). Using a causal learning paradigm in which participants had to learn about the causal strength between two types of events, Danks et al. (in press) showed that while people's blame judgments were influenced by the moral nature of the events, their judgments of causal strength were not affected. They take this as evidence that the strong hypothesis about an intimate linkage between causality and morality is mistaken: some aspects of causal cognition are immune to evaluative differences. Future research will need to identify when morality affects causality and when not (see Cushman & Young, 2011; Guglielmo, Monroe, & Malle, 2009).

2.2.3 Strengths and weaknesses of label theories

I have provided an overview of two sets of attribution theories which both fall under the header of *label theories*: first, normative theories which prescribe how an ideal observer should attribute causality, responsibility and blame (Shaver, 1985; Shultz et al., 1981, e.g.) and, second, descriptive theories which aim to characterise how people actually arrive at their attributions of responsibility and causality (e.g. Alicke, 2000; Knobe, 2010). While both normative and descriptive theories of responsibility attribution have picked out similar factors, such as causal control or intentions, the two families of theories differ markedly in the way in which they conceive of the relationships between these factors. Normative theories conceptualise the relationship between causality, responsibility and blame in terms of entailment (i.e. causality $\rightarrow$ responsibility $\rightarrow$ blame). Descriptive theories, in contrast, have provided empirical evidence that these different concepts are closely linked with each other (Knobe, 2010) and that moral evaluations can influence judgments of causality (Alicke, 2000).

Label theories of attribution have both strengths and weaknesses. On the positive side, they have identified a multitude of factors that influence people's attributions. A great deal of empirical work has subsequently shown that people are indeed sensitive to these factors and that attributions of responsibility increase (e.g. if the outcome was foreseen and intended, Lagnado & Channon, 2008) or decrease (e.g. if mitigating circumstances such as situational or capacity constraints were present, Alicke, 2000; Woolfolk, Doris, & Darley, 2006) as predicted. Furthermore, label theories (whether normative or descriptive) have attempted to provide criteria for distinguishing between related concepts such as causality, responsibility, blame and punishment (Cushman, 2008; Fincham & Jaspars, 1980; Fishbein & Ajzen, 1973; Hart, 2008; Robbennolt, 2000; Shaver & Drown, 1986; Shultz & Schleifer, 1983). Following the suggestion by Shaver and Drown (1986), many researchers have used multiple response measures (e.g.

separate scales for *cause*, *responsibility* and *blame*) to enable participants to express distinctions between the concepts. The success of this strategy has been mixed: while some studies find patterns of responses that differentiate between the concepts (Harvey & Rule, 1978; Lagnado & Channon, 2008; Shaw & Sulzer, 1964; Shultz et al., 1981; Tyler & Devinitz, 1981; Weiner, 1995) the different dependent variables tend to be quite highly correlated (Critchlow, 1985) and are sometimes not significantly different from each other (Alicke, Buckingham, Zell, & Davis, 2008).

The notion of responsibility is particularly difficult to pin down. Many researchers have acknowledged that 'responsibility' is an inherently polysemous term (e.g. Gailey & Falk, 2008; Hart, 2008; Sartorio, 2007; Schlenker, Britt, Pennington, Murphy, & Doherty, 1994; Sousa, 2009; Zimmerman, 2001). It can be used to refer to a dispositional attribution ("Suzy is a responsible person"), to a specific role or obligation ("It is your responsibility to wash the dishes"), to an action ("Tim behaved irresponsibly") or to an outcome ("John is responsible for our team's loss").[13] Furthermore, one can distinguish *causal* from *moral* and *legal* responsibility (Fincham & Jaspars, 1980; Hart, 2008; Robbennolt, 2000).

Schlenker et al. (1994) argue that there are two core facets of responsibility: the first facet is closely tied to causality (they use the term *imputation*) and the second facet is related to the idea of accountability. While the first facet captures who did it and why, the second facet relates to whether the actor is liable to social sanctions (Fincham & Jaspars, 1980; Hamilton, 1978). Both Hart (2008) and Hilton and Slugoski (1986) have argued that the scientific concept of causality on which some of the attribution models are based does not fully capture the richness of responsibility attributions. Hart (2008) even went so far as to explicitly exclude a scientific concept of causality from his analysis of responsibility. Related to the concept of accountability, Hamilton (1978, 1986) introduced the idea that responsibility is relative to the obligation of a person that comes with their role. For example, if we observe that a policeman and a fireman do not help a stranger on the street who is in trouble, we might attribute more responsibility to the policeman for the failure to act because it is part of his role to intervene in such circumstances. This conception of responsibility is closely related to the notion of vicarious responsibility as discussed above. A military officer is responsible for the actions of her subordinates as part of her social role. In Chapter 4 we will see that people's responsibility attributions are indeed sensitive to the different roles that individuals have in a particular situation. As we will see below, most formal theories of attribution have neglected these conceptual differences.

One of the major weaknesses of label theories is a very general one: theories that don't specify their concepts in terms of formal models necessarily lack precision. For example, Darley and Shultz (1990) argue that "in cases where it is judged that harm

[13] In the United Kingdom, advertisement for alcoholic drinks is always accompanied by the cautionary advice: "Please drink responsibly." Unfortunately, the advice appears to be overlooked quite frequently.

was done intentionally, blame is *a joint function* of moral responsibility, the presence of net harm, and justification for the harm." (p. 533, emphasis added) However, they do not spell out what this *joint function* is supposed to look like. They are neither specifying what the exact relationships between the different concepts are nor how strongly different factors (e.g. intention vs. outcome severity) are predicted to influence blame attributions. For example, intentionally committed actions that lead to negative outcome are deemed *more* blameworthy than accidental actions. However, how much more blameworthy? Given the imprecise nature of label theories, it can be difficult to devise experiments to falsify their predictions. In contrast, a formal theory which makes precise predictions can be readily tested empirically and refined in case the results are at odds with its predictions.

Furthermore, almost all of the empirical evidence that supports label theories comes from a single methodological approach: vignette studies. Participants are generally asked to judge the behaviour of a protagonist in a scenario, based on the information they are provided with (usually, different factors are merely manipulated qualitatively, i.e. the protagonist either intended the outcome or not). Especially the burgeoning field of experimental philosophy is subject to this criticism: there hasn't been a single study in this field thus far that has not relied on scenario-based experiments. In part this is surely due to the fact a philosopher's curriculum generally does not feature instructions about experimental techniques. Often, however, the problems with vignette studies are not sufficiently acknowledged and bold theoretical claims are drawn from relatively weak empirical evidence (see Huebner, 2011; Scholl, 2008, for related criticisms). Researchers often employ quite unrealistic scenarios that lack ecological validity and it is unclear whether evidence about people's intuitions garnered from these cases generalises to other situations.

What is more, scenarios are necessarily incomplete descriptions of the actual circumstances and participants are required to fill in the gaps. This is an obvious source of interindividual variability as different people are likely to fill in the gaps in different ways. Just to mention one example: while participants might be convinced by the description of the scenario that pushing the fat man in one version of the famous trolley dilemma (Thomson, 1976) is really the only way to stop the train from killing the five people lingering about on the track, two participants could still differ in their assumptions about the probability that their intervention will be successful. A participant who thinks that there is a possibility that pushing the fat guy will actually not stop the train, will reach a different moral judgment than another participant who does not question the potential fallibility of this intervention. The researcher, however, has no handle on finding out how participants interpreted the scenario (apart from asking for explanations, hoping that participants will provide them *and* being able to code and analyse them).

2.3 Formal Theories of Responsibility Attribution

In this section, I will give an overview of different formal theories of responsibility attribution. These theories differ from the label theories discussed in the previous section in that they make precise, quantitative predictions about how certain factors influence attributions. Figure 2.8 shows some of the formal tools that researchers have drawn upon to develop models of causal and responsibility attribution. These different tools constitute a diversity of languages for expressing people's attributional capacities. I will argue that some languages are better suited for this task than others. A formal framework for responsibility attribution needs to be grounded in a language that allows to represent the causal structure of the world. This representation should not only support inferences based on what happened in the actual world but also on what could have happened in other possible worlds. Since most theories have not relied exclusively on one type of tool but have used a combination of different tools, I will categorise different theories into broader categories.

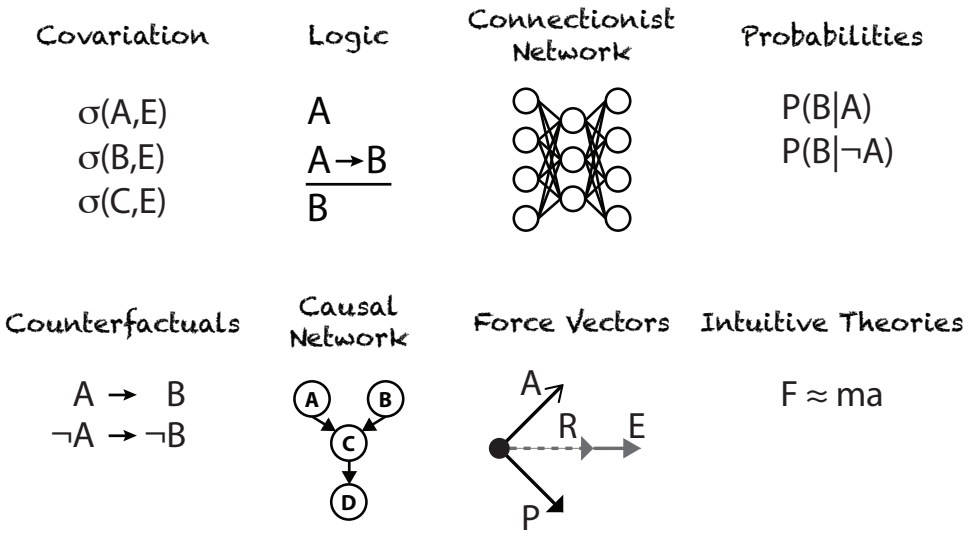


Figure 2.8: Formal tools for attribution theories.

2.3.1 Covariation, logic and connectionism

In the introduction to this chapter, I have introduced the work of Kelley (1967, 1973), which can be seen as the first formal treatment of the attribution process. Kelley's (1973) ANOVA model likens the attribution process of the layperson to the formal statistical way in which scientists draw conclusions from their experimental procedures. The principle of *covariation* serves as the formal tool to model people's attributions. An event A is held causally responsible for another event B, if A and B covary over time. That is, whenever A is absent B is absent and whenever A is present B is present. We

have already noted some of the problems with the covariational account above. Most importantly, based on covariation alone, it is impossible to distinguish causes from effects (Cheng, 1997; Glymour, 2007; Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011; Jenkins & Ward, 1965; Meder, Gerstenberg, Hagmayer, & Waldmann, 2010; Shanks, 1995; Waldmann & Hagmayer, 2005; Waldmann & Holyoak, 1992). If all we know about A and B is that they covary over time, it could be that A causes B or that B causes A (whenever I see my brother's shadow, my brother happens to be there – but surely, my brother causes his shadow and not vice versa). What is more, A and B could both be effects of a common cause C. However, we would not want to hold a person's yellow fingers responsible for his lung cancer. Instead, it is the fact that he smokes which causes both the colour of the fingers and the state of his lung to covary over time.

Having the formal tools to distinguish between causes and effects is fundamental to a framework of responsibility attribution. While covariational accounts cannot make this distinction, accounts based on *logic* can capture an asymmetry between two events via the relationship of material implication. If we know that A is present and B happens whenever A happens ($A \rightarrow B$), we can deduce the presence of B. From the presence of B and knowing that $A \rightarrow B$, however, we cannot deduce that A was present. Intuitively, B could have been brought about by another event. In other words, the material implication $A \rightarrow B$ implies that A is sufficient (but not necessary) for B, whereas B is necessary (but not sufficient) for A. However, a substantial amount of psychological research has shown that logic-based accounts are incapable of expressing people's causal inferences (Oaksford & Chater, 2007; Sloman, 2005; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). First, logic-based accounts only work in the realm of certainty in which there is no doubt about whether a particular condition was met. However, our world is fundamentally uncertain and most of our inferences are based on uncertain pieces of evidence (see, e.g. Oaksford & Chater, 2007). Second, logic-based accounts cannot distinguish inferences based on observations from inferences based on interventions (cf. Glymour, 2007; Sloman, 2005; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). We will see in Chapter 3 that this capacity is crucial for an adequate theory of responsibility attribution.

Notwithstanding these limitations, there have been some attempts to develop logic-based computational models of responsibility attribution (e.g. Mao & Gratch, 2005; Shultz, 1986). Darley and Shultz (1990, p. 538) give an example of such an account which conceptualised attributions of blame in terms of a production system with 39 *if-then* rules (Shultz, 1986): "If the protagonist produced the harm and the protagonist's action was not accidental and the protagonist's action was voluntary and the harm was a foreseeable consequence of the protagonist's action and there was no intervening cause of the harm, then the protagonist is morally responsible for the harm." In a similar vein, Mao and Gratch (2005) took Shaver's (1985) model as a starting point and incorporated

the different factors, such as foreseeability and intentionality, as logical conditions into a model of responsibility attribution.

There are obvious problems with such logic-based accounts. First, the different conditions need to be defined as either absent or present. Often, however, the outcome of an action is neither *foreseeable* nor *not foreseeable* but *foreseeable to a certain degree*. The fact that the different attributes don't come in a neat binary packaging makes it problematic for logic-based accounts to determine which *if-then* rule (if any) is applicable in a given situation. Second, most of the theoretically interesting questions are already incorporated into the rules. For example, what does it mean in Shultz's (1986) model for a person to have *produced* the harm? Ideally, we would like a model that determines the degree to which a person produced the harm as a function of the different events that happened and the ways in which they were causally related. We will see in Chapters 3 and 6 that a different framework allows us to answer questions about which out of several candidate events have actually caused the outcome. Finally, logic-based accounts have notorious problems with exceptions. For example, if the protagonist above suffered from insanity, it would still be correct to say that his action was voluntary but we might be less inclined to say that the person was morally responsible for the produced harm.

Darley and Shultz (1990) argue that connectionist models could be another promising formal framework for modelling attributions of responsibility. Connectionist architectures can yield graded responses and have the advantage of being able to cope with noisy stimuli (Rumelhart & McClelland, 1988). Thus far, however, no connectionist model of responsibility attribution has been developed. Connectionist architectures are based on an associative machinery which does not allow to distinguish between causes and effects. Since connectionist models lack the capacity to represent the causal structure of the world, I would argue that they are non-starters as models of responsibility attribution.

2.3.2 Probabilities and counterfactuals

The majority of formal models of responsibility attribution developed thus far has relied on the formal tools of probabilities and counterfactuals. I will first discuss an exclusively probabilistic account that has focused mostly on dispositional attribution. I will then describe accounts that model responsibility attributions in terms of counterfactuals.

2.3.2.1 Ajzen and Fishbein's (1975) Bayesian account of causal attribution

Inspired by Kelley (1967, 1973) and Jones and Davis (1965), Ajzen and Fishbein (1975, 1983) provide a unifying formal framework for many of the factors that have been found to influence causal attributions, such as consistency of behaviour across situations, objects and actors (see Kelley, 1973), and the actor's perceived decision freedom (see Jones & Davis, 1965). Ajzen and Fishbein (1975, p. 265) state that "attribution theory

deals with the perceived likelihood of alternative causal factors as explanations of observed behavior". They propose that people's causal attributions can be conceptualised in terms of belief revision: people tend to attribute an observed behaviour to the causal factor in whose light the behaviour was most likely to occur (compared to other possible factors that could have also caused the behaviour).

Equating attribution with belief revision opens the door to a Bayesian analysis of the attribution process (see Peterson & Beach, 1967, for a review of early work on statistical inference models of cognition). Bayes' theorem states how beliefs about different hypotheses (expressed as subjective probabilities) are to be revised in the light of observed evidence. The subjective degree of belief about a particular hypothesis H_1 in the light of some observed data D should be revised in the following way:

$$p(H_1|D) = \frac{p(D|H_1) \times p(H_1)}{\sum_{i=1}^{n} p(D|H_i) \times p(H_i)}, \tag{2.1}$$

whereby $p(H_1|D)$ is the posterior probability of hypothesis H_1 given the observed data D, $p(H_i)$ is the prior probability of hypothesis i, $p(D|H_i)$ is the likelihood of the observed data D given hypothesis H_i and n equals the number of hypotheses.

Consider a situation in which Ted makes the following statement (D_1) to his wife as they are about to leave for a dinner party: "Gosh, you look beautiful tonight!" Several hypotheses about causal factors that might have brought about this behaviour come to mind: (1) Ted just remembered how beautiful his wife was, (2) Ted is in a somewhat euphoric state because he is looking forward to the dinner party, (3) Ted wants to please his wife because he anticipates that his friend's dinner party will not be to her liking. I will use the labels H_{love}, $H_{euphoria}$, $H_{strategy}$ for the three different hypotheses. Let us assume, for simplicity, that these three hypothesis are mutually exclusive and exhaustive. That is, only one of the three hypothesis can account for the observed behaviour and there are no other possible explanations. In reality, it is of course often the case that several causal factors are collectively responsible for an observed behaviour.

How should an ideal observer update her beliefs about what has driven Ted to his remark? First, we need to assign a prior probability H_i to the different hypotheses. These priors will depend on how well the observer knows Ted and in what circumstances he has interacted with Ted thus far. For simplicity, we will assume that each of the hypotheses is equally likely a priori. Next, we need to determine $p(D|H_i)$, the likelihood of the particular behaviour (i.e. the paid compliment) under the different hypotheses. Ajzen and Fishbein (1975) argue that the likelihood ratio of each hypothesis is of particular interest for the attribution process. The likelihood ratio (LR) refers to the ratio between the probability of the behaviour under a given hypothesis $p(D|H_1)$ and the probability of that same behaviour given all other hypotheses $p(D|\neg H_1)$, hence, $LR_{H_1} = \frac{p(D|H_1)}{p(D|\neg H_1)}$. The greater the likelihood ratio for a particular hypothesis, that is, the more likely a

hypothesis becomes in the light of the observed behaviour compared to the other hypotheses, the more likely people should see this hypothesis as responsible for having caused the observed data (see Ajzen, 1971, for a direct empirical test of these model predictions).

Let's suppose that the observed behaviour thus far is equally likely to occur under the hypotheses H_{love}, $H_{euphoria}$ and $H_{strategy}$. Hence, based on this single piece of evidence, we cannot distinguish between the different hypotheses. An important feature of Bayes' rule is that there is a straightforward way of sequentially updating hypotheses as additional evidence comes in ('today's posterior is tomorrow's prior'). Let's assume that on the way to the party, Ted makes another comment (D_2): "Don't you also have the feeling, honey, that everything looks so beautiful today? The trees, the streets the evening sky, ...". Given this observation, Bayes' rule predicts that our belief about $H_{euphoria}$ will increase and the belief in the other hypotheses correspondingly decrease (i.e. $LR_{H_{euphoria}}$ is greater than $LR_{H_{love}}$ and $LR_{H_{strategy}}$). We might now be fairly confident that Ted is just in a very good mood tonight which explains both pieces of evidence D_1 and D_2. Our inference would have been different, if D_2 had been "Wow, my hands are still quite sore from all the cleaning that I have done in the house today" or a spontaneous kiss in the car. The Bayesian account also predicts that people's causal attributions should be sensitive to the sample size. Thus, the more data is observed that is congruent with a particular hypothesis, the more the posterior belief in this hypothesis will increase and the more confident one should be about the underlying causal factor that gives rise to the data.

Ajzen and Fishbein (1975) discuss that much of the empirical work that has looked at people's attributions can be interpreted as having manipulated the diagnosticity of the evidence for different hypotheses. For example, when a person is coerced to show a particular behaviour, that behaviour is not diagnostic any more about the person's actual dispositions (Jones & Davis, 1965). However, when the person acts in the absence of external constraints, an inference about the person's desires is warranted. The Bayesian analysis shows how different patterns of inference such as the *covariation principle* (belief revision in light of repeated evidence) and the *discounting principle* (mutual dependence of hypotheses) can be unified into one coherent framework (see also Morris & Larrick, 1995).

2.3.2.2 Brewer's (1977) information-processing model

The first formal model which has explicitly addressed the problem of responsibility attribution has been developed by Brewer (1977). Her model is based on two formal tools: counterfactuals and probabilities. We will hear much more about counterfactuals in Chapter 3 so I will only provide a short sketch here. According to a counterfactual theory of causality, two conditions have to be met in order for an event C to qualify

as a cause of another event E. First, both C and E must have occurred (i.e. the *actuality condition*) and, second, it must be true that if C had not occurred then E would not have occurred (i.e. the *counterfactual condition*). Applied to responsibility, we should say that a person's action C is responsible for an outcome E, if he acted and the outcome occurred, *and*, if the outcome would not have occurred *but for* his action.[14] Hence, counterfactual theories of responsibility attribution predict that a person's responsibility is not only determined by what actually happened but also by what would have happened had the person acted differently.

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. Now consider two possible versions of the story: in story one, Bill and Suzy dine at *Restaurant contrôle effective*. In this restaurant, it is the case that if Bill had ordered meal B, Suzy would have been just fine. 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 your judgment is that the causal significance of Bill's choice is greater in *Restaurant contrôle effective*, you are in line with what most participants indicated. 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. Nevertheless, as this result shows, people's attributions were markedly influenced by counterfactual considerations. If the same outcome would have prevailed no matter what Bill did, he is not seen as very responsible. If, in contrast, Bill could have prevented the negative outcome had he acted differently, he is seen as responsible for the outcome.[15]

The scenario above was set up in a way that left no uncertainty about what would have happened in the relevant counterfactual worlds. However, the real world is noisy and we cannot observe counterfactual outcomes. This is where the probabilistic part comes in: according to Brewer's (1977) model, people's responsibility judgments involve the comparison of one's subjective degree of belief that the effect will occur in the

[14]The same *but for* test is used as a definition of *factual* causation in the law (Hart & Honoré, 1959/1985; Spellman & Kincannon, 2001). Additional criteria, such as sufficient proximity between cause and effect, have to be met to qualify as a *legal* cause.

[15]Wells and Gavanski (1989) only asked for causal judgments and not for responsibility judgments in the restaurant scenario. However, they asked for both causal and responsibility judgments in Experiment 2 which conceptually replicated the restaurant scenario.

presence of the cause $P(E|C)$ with one's belief about whether the effect would have occurred in the absence of the cause $P(E|\neg C)$. More specifically, $P(E|C)$ "refers to the extent to which the outcome would be likely, *a priori*, to follow from the action, as perceived after the fact" (Brewer, 1977, p. 59, emphasis in original).[16]

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 component is additive (see Equation 2.2). 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.

$$Responsibility(C) = P(E|C) - P(E|\neg C) \tag{2.2}$$

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 for the restaurant episode mentioned above. Here, the probability of Bill's death given Sarah's shot, $P(E = \textit{Bill dies}|C = \textit{Sarah shoots})$, is high. Furthermore, under most circumstances we can expect that the probability of Bill's death in the absence of Sarah's shot, $P(E = \textit{Bill dies}|\neg C = \textit{Sarah doesn't shoot})$, is very low. Thus, Sarah is predicted to be judged highly responsible for Bill's death. In short, according to Brewer (1977), attributions of responsibility are related to the degree to which a person's action made a difference to the outcome, while *difference* is measured in terms of how much the observer's subjective estimate of the outcome's probability has changed in light of the person's action.

Brewer (1977) discusses that her model accounts for how an actor's motive or the severity of the outcome affect responsibility attributions. Brewer's (1977) model predicts that the actor's motives (such as her intentions) should only matter to the degree that they affect the observer's belief about $P(E|C)$ and $P(E|\neg C)$. Shaw and Sulzer (1964) have found that judgments of responsibility increase with the foreseeability of the outcome even when the actor's intention is controlled for (see also Lagnado & Channon, 2008). Hence, in situations in which there was no ambiguity about the relationship between a person's action and the outcome, knowing that the person intended this outcome should not affect attributions of responsibility. In contrast, when the situation was such that there were many different effects that could have resulted from an action, knowing that the person intended this particular outcome reduces the uncertainty about $P(E|C)$ and should accordingly lead to an increase in attributed responsibility.

[16]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 (cf. Woodward, 2006). The difference between conditioning based on observation versus counterfactuals will be discussed in more detail in Chapters 3 and 5.

In general, research has shown that there is a positive relationship between outcome severity and attributions of responsibility although results have been somewhat mixed (see Robbennolt, 2000). Brewer (1977) argues that because the probability of a severe outcome is normally quite low, the influence of a person's action on the probability of severe outcomes is *potentially* very high. However, her model predicts that people would only assign responsibility when the probability of the severe outcome was significantly increased due to the person's action. In many situations, the probability that a severe negative outcome follows from one's action is also low.

For example, imagine that Gary decided to drive back home with his car despite the fact that he was sleep-deprived. On his way home, he crashes into a crash barrier and causes some significant damage. According to Brewer's (1977) model, Gary should be judged highly responsible for this outcome. The chances that a mild negative outcome such as crashing into a crash barrier or hitting a tree would result from driving when being tired are high.

Now contrast this scenario with one in which the outcome is much more severe: Gary did not hit the crash barrier but his car happened to come off the road, came to a halt on a railway track, caused an oncoming train to derail and led to the death of many people. According to Brewer's (1977) model, Gary should not be held very responsible for this severe negative outcome.[17] The chances that something like this would happen given that he drove back home being tired are very slim. Hence, the model is sensitive to what philosophers have called moral luck (Nagel, 1979; Williams, 1981): the degree to which a person is judged responsible is *not* contingent on the severity of the outcome *per se* but only on the probability that a particular outcome could be expected to arise from the person's action (and the probability that this outcome was likely to occur in the absence of the person's action). We will learn more about moral luck and the relationship between intentions and outcomes in Chapter 5.

2.3.2.3 Fincham and Jaspars's (1983) subjective probability model

Fincham and Jaspars (1983) take Brewer's (1977) model as a starting point and argue that it lacks a crucial aspect that moderates attributions of responsibility. As Fincham and Jaspars argue, Brewer's model can be interpreted as predicting responsibility attributions based on internal factors (a person's action) and external factors (the situation as it would have been in the absence of the person's actions). The missing piece from Kelley's analysis concerns the relationship between the situation and the action. Different situations carry different normative forces and attributions of responsibility are particularly attracted by behaviour that is counter to the observer's expectations (cf.

[17] This example is based on a true story. The Great Heck rail crash (also known as the Selby rail crash) is the worst rail disaster of the 21st century in the UK: 10 people died and 82 suffered serious injuries. Gary Hart was sentenced to five years imprisonment for causing death by dangerous driving (Wikipedia, 2012).

Jones & Davis, 1965).

Thus, Fincham and Jaspars (1983) propose that responsibility attributions are not only sensitive to the two components identified by Brewer (1977), $P(E|C)$ and $P(E|\neg C)$, but also to the subjective degree of belief that another person would have acted the same way in the given circumstances (cf. Kelley's, 1973, consensus dimension). Fincham and Jaspars (1983) term the probability that another person would have acted the same in a given situation *validation*. Validation is negatively related to attributions of responsibility. That is, the higher a person's degree of belief that another person would have acted the same way in a given situation, the less will the person be seen as responsible.

Fincham and Jaspars (1983) criticise that Brewer's (1977) model does not distinguish between causation and blame (or causal responsibility vs. blameworthiness). However, as discussed above, there is evidence that people's attribution of blame and causation can go apart to some degree (see, e.g. Fincham & Shultz, 1981; Lagnado & Channon, 2008; Shultz et al., 1981). Fincham and Jaspars (1983) argue that their notion of *validation* can serve to distinguish causation from blame: while conceptualising the degree to which a person's action made a difference to the outcome in terms of the comparison between $P(E|C)$ and $P(E|\neg C)$ might be sufficient to capture *causal* attributions, whether the person will be *blamed* or not crucially depends on validation. Fincham and Jaspars (1983) report five experiments which broadly support their account. In general, attributions of causality, responsibility and blame increased with $P(E|C)$ as predicted by Brewer's (1977) model: however, responsibility and blame attributions (but not causal attributions) were moderated by validation. The protagonist was judged more responsible and blamed more if the participant thought that other people would have acted differently in the given situation. In contrast, participants' validation judgments did not affect their causal attributions.

2.3.2.4 Spellman's (1997) crediting causality model

Similar to Fincham and Jaspars (1983), Spellman (1997) has also taken Brewer's (1977) model as a starting point. However, rather than considering the influence of *validation*, Spellman (1997) stuck to the core of Brewer's (1977) model but extended it in order to handle situations in which an outcome was brought about by a sequence of causal events. For example, consider the following scenario (Miller & Gunasegaram, 1990):

> 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 more 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.

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* the causal event has occurred with the probability of the outcome *after* the event. In line with Brewer (1977), 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 2.3). 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 $P(win|Jones\ tossed\ head) = 50\%$. For Cooper, in contrast, the situation looks different. Before he tosses his coin, the probability of them winning is 50%. However, after he tossed his coin, the probability either increases to 100% if the coins match or it decreases to 0% if the coins mismatch. Hence, while Jones's action does not change the probability of the outcome, Cooper's does.

$$\begin{aligned} Responsibility(Jones) &= P(win|Jones\ tossed\ head) - P(win) \\ &= 0.5 - 0.5 = \mathbf{0} \\ Responsibility(Cooper) &= P(win|Cooper\ tossed\ head\ \&\ Jones\ tossed\ head) \\ &\quad - P(win|Jones\ tossed\ head) \\ &= 1 - 0.5 = \mathbf{0.5} \end{aligned} \tag{2.3}$$

In a series of experiments, Spellman (1997) varied the extent to which the probability of the effect was changed by different causal events in a chain. The crediting causality model predicted people's causal judgments as well as their attributions of blame very accurately. 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 model 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, imagine that Bill 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 if it is the case 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. 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 (see 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 an 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 favourite 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 is walking 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 almost 100%. However, in a similar scenario that Mandel (2003) used in his experiment, 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. While being poisoned increased Bill's probability of dying, it did not increase the chances of him dying *from a gun shot*. However, we will see in Chapter 3 that the strategy of increasing the granularity of the effect will not always help.

2.3.2.5 Petrocelli, Percy, Sherman, and Tormala's (2011) counterfactual potency model

Let us conclude this review of theoretical accounts that propose a close relationship between counterfactuals, probabilities and attributions of responsibility with a recent addition. According to Petrocelli et al. (2011), attributions of responsibility are related to the *potency* of a relevant counterfactual. We have seen above that counterfactuals

can be expressed in *if ... then* form. A counterfactual's potency is a function of (i) the likelihood that the alternative considered antecedent could have come about (i.e. the *if*-part) and (ii) the likelihood with which the alternative antecedent would have undone the outcome (i.e. the *then*-part). A counterfactual's potency is a multiplicative function of the *if-likelihood* and the *then-likelihood* (see Equation 2.4).

$$\textit{Counterfactual Potency} = \textit{if-likelihood} \times \textit{then-likelihood} \tag{2.4}$$

Let me illustrate the model predictions 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*. 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* how likely it was 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.

Research in social psychology has identified a host of factors that influence when counterfactual thoughts are elicited and what aspects of a situation are counterfactually 'manipulated'. People tend to consider counterfactuals to events that were abnormal, controllable or close to a desired outcome (e.g Byrne, 2002; Girotto, Legrenzi, & Rizzo, 1991; Kahneman & Miller, 1986; Roese, 1997). In Petrocelli et al.'s (2011) terms, these factors are what influences the subjective *if-likelihood* of the counterfactual. The degree to which the different counterfactual thoughts influence attributions of responsibility now depends on how effective they are imagined to be in undoing the outcome through mentally simulating what would have happened (see Kahneman & Tversky, 1982).[18]

Petrocelli et al. (2011) demonstrate in a range of studies that their model accurately predicts the qualitative trends in participants' attributions. In particular, their studies support the assumption that *if-likelihood* and *then-likelihood* combine in a multiplicative

[18]The importance of mental simulation for causal and responsibility attributions will be discussed in much greater detail in Chapter 6.

fashion. Hence, attributions of responsibility are high only if *both* the *if-likelihood* and *then-likelihood* of the relevant counterfactual are high.

Since Petrocelli et al.'s (2011) model is fairly novel, there have not been any theoretical or empirical attempts to point out limitations of their model. However, I see several ways in which their account is limited in important respects (at least as a model of responsibility). Let me just briefly make one point here:

While I have focused in my discussion of their model on the relationship between *counterfactual potency* and *responsibility*, their account is actually proposed to be much more general. For example, it has also been used to predict how people make attributions about how much *regret* a person will experience in different situations (Experiments 1, 2 and 4 of their paper). That there is a relationship between counterfactual potency and experienced regret is very intuitive. Let us consider the restaurant scenario above once more. The intuition is strong that Bill would feel more regret for his choice in *Restaurant contrôle effective* in which the alternative choice would have undone the negative outcome. Furthermore, as the model predicts, experienced regret should only be high when he considered going for the alternative option.

In their Experiment 2 (which is conceptually identical to the dinner scenario), Petrocelli et al. (2011) find the predicted effects of counterfactual potency on *both* judgments of regret and responsibility/blame. However, let me now slightly tweak the scenario to show that the relationship between counterfactual potency, regret and responsibility is not quite as straightforward. Let's imagine that, this time, Bill has the intention to kill Suzy and takes her to *Restaurant contrôle effective* (which is just around the corner from where Bill lives). Bill knows of Suzy's allergy and foresees that if she will eat meal A, Suzy will die for sure. Without thinking twice, Bill orders meal A, Suzy eats it and dies. If now being asked to say how much regret Bill is likely to experience for the outcome, the counterfactual potency model gets it right. Given that he got exactly what he wanted, Bill will presumably not regret what he did. However, if we now consider how responsible Bill is for Suzy's death, the model gets it wrong. Most models will say that the fact that Bill intentionally killed Suzy in a way just as foreseen makes him very responsible and blameworthy for the outcome (e.g. Shaver, 1985). However, the *if-likelihood* of an alternative action in the scenario is very low and thus the counterfactual potency model predicts that Bill's responsibility for the outcome should be low.

It is easy to create the reverse scenario for which the intuition is that experienced regret will be high but responsibility low (or at least lower) while the counterfactual potency model predicts that *both* regret and responsibility should be high. If Bill had been in two minds while making the order and questioned whether he really wanted to go ahead with the plan as intended then this would increase the *if-likelihood* of the relevant counterfactual. Indeed, in this version of the scenario it seems natural that Bill might experience a lot of regret for what he did (assuming he ended up choosing meal A in the end). However, the fact that his intention in this situation was weaker and he

was unsure about whether to pursue his evil plan or not will arguably make him *less* responsible (cf. Holton, 2009, for a philosophical discussion of the relationship between intentions and weakness of will).

Together these two versions of the scenarios show that, at least as a model of responsibility, Petrocelli et al.'s (2011) model is limited. While the other formal models discussed above correctly predict that intentions and foreseeability increase attributions of responsibility (at least to the extent that these mental states influence the probability of the outcome given the person's action), the counterfactual potency model gets it the wrong way round.

2.3.2.6 Bayesian networks, force vectors and intuitive theories

The models discussed in the previous section have argued for a close correspondence between responsibility attributions and counterfactual considerations. The extent to which a person's action will be deemed responsible for an outcome depends on how likely the outcome would have been undone had the person acted differently. However, these models have failed to provide a language that captures how people represent the causal relationships on which the evaluation of counterfactuals is based. These accounts only stipulate that an attributor's belief about what would have happened in a relevant counterfactual world will influence their attributions. They have been silent, however, as to how people arrive at this belief.[19] Without a representation of the causal structure that supports the consideration of counterfactuals, these models provide at best a partial account of how people reach their responsibility judgments.

Let me just briefly illustrate this point via the Petrocelli et al. (2011) model. Their model predicts that when an event has a potent counterfactual (i.e. a counterfactual with a high *if-* and *then-likelihood*), it will be held responsible for the outcome. However, their model is not based on a causal representation of the situation. Hence, it would predict that in a common cause structure ($E_1 \leftarrow C \rightarrow E_2$), one effect E_1 will be judged responsible for another effect E_2. Assuming that there is a deterministic relationship between the cause and the two effects, the counterfactual if E_1 had been different then E_2 would have been different might well be true and potent assuming that C is likely to happen.[20] However, surely we would not want to hold one effect responsible for another effect – instead, C is responsible for the occurrence of both E_1 and E_2. In recent years, new formal languages have been developed to represent people's causal beliefs about the world.

Kelley (1983) has argued that people "perceive the temporal flow of life's events to be structured causally" (p. 343) and that we understand particular causal events in

[19] In many experiments, participants were explicitly informed about what the chances of the outcome would be under different possible contingencies (cf. Spellman, 1997).

[20] Whether or not the counterfactual statement is true depends on exactly how counterfactuals are operationalised (cf. Hiddleston, 2005; Rips, 2010).

terms of how they are embedded within a greater causal structure. Basic properties of such causal structures are (i) that they reflect the temporal order of the how the events unfolded in the world, (ii) that different causal events can be connected in complex ways (one cause may have many effects and/or one effect many causes), (iii) that causes differ in terms of their proximity to the effect, (iv) that causal relationships differ in terms of their stability and (v) that causal structures support both inferences about what actually happened as well as about what could have happened. Kelley (1983) notes that early research in attribution theory neglected these complexities in favour of simple explanations: "It was once assumed that internal and external (or disposition and situation) attribution ratings would be inversely related, inasmuch as these terms seemed to represent opposite poles of explanation. The conception of perceived causal structures, in which effects are often multiply determined, makes this assumption seem quaintly simple." (p. 358)

The formal framework of causal Bayesian networks (Pearl, 1988, 2000; Sloman, 2005; Spirtes, Glymour, & Scheines, 2000) incorporates the features that Kelley (1983) saw as core properties of how people represent the world. A causal Bayesian network (CBN) can represent a system of an arbitrary number of causes that are linked via complex causal relationships. CBNs can be interpreted as normative models that prescribe how a person should update their beliefs about causal events within a given causal structure based on evidence garnered from observing parts of the system and actively intervening in it.[21] Finally, since a CBN represents the causal structure of the world, it can also be used to reason about hypothetical or counterfactual interventions. That is, a CBN supports predictions about what would happen if one was to perform a certain action or about what would have happened if a certain event had turned out differently. In that sense, CBNs predict a close relationship between causal reasoning and imagination (see Lagnado, 2011a; Walker & Gopnik, forthcoming, for arguments that people's simulation of possible worlds respect the causal structure of the actual world).

Recent work in causal attribution has built heavily upon this framework. For example, Baker, Saxe, and Tenenbaum (2009) have shown that people's inferences about an agent's beliefs and desires can be explained in terms of inverse planning. That is, assuming that an agent plans their actions rationally based on their beliefs and desires about the world, we can use Bayesian inference to invert that process and infer an agent's beliefs and desires from the actions (Baker, Saxe, & Tenenbaum, 2011). Using experimental stimuli in which an agent navigates an environment, these models provide good quantitative fits to people's inferences about the agent's goals (Goodman, Baker, & Tenenbaum, 2009) or even joint inferences about both the agent's goals and beliefs (Baker et al., 2011). Similar models have also shown that children's performance in experiments can be well-described in terms of rational inferences about other agents'

[21] See Cartwright (1995) for a criticism of the CBN approach.

beliefs and desires (Richardson, Baker, Tenenbaum, & Saxe, 2012) or the properties of objects (Gweon & Schulz, 2011; Gweon, Tenenbaum, & Schulz, 2010).[22]

However, up until recently, the potential of exploring CBNs as a representational framework for understanding how people attribute responsibility has not been utilised (McCoy, Ullman, Stuhlmüller, Gerstenberg, & Tenenbaum, 2012). In Chapter 3, I will show how causal Bayesian networks work and how they can be used to derive predictions about how people attribute responsibility to a particular event within a causal structure. I will propose that people's responsibility attributions can be modelled in terms of counterfactuals defined over a causal representation of the situation. This proposal essentially combines insights from the counterfactual models of responsibility attribution discussed in the previous section with a richer formal language to express people's causal beliefs about the world. Subsequent chapters will then empirically test predictions derived from this framework.

In Chapter 6 we will discover some of the limitations of the CBN approach. People's causal understanding of certain domains surpasses the expressive power of the CBN framework (Tenenbaum, Griffiths, & Niyogi, 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). For example, CBNs have difficulty in representing information about continuous space and time which is crucial for capturing people's understanding of the physical world. These limitations have led some researchers to adopt a very different approach of modelling people's causal understanding of the world (Wolff, 2007). Accordingly, people are assumed to represent the world in terms of interacting forces rather than in terms of abstract counterfactual dependencies between events as incorporated in a CBN. However, I will show that both ways of modelling attributions can be unified. People's attributions are informed by counterfactuals but these counterfactuals need not be defined in terms of CBNs. Rather, people use their intuitive domain theories (which are often richer than what can be expressed in CBNs) to arrive at their beliefs about what would have happened in different possible worlds.

2.3.3 Strengths and weaknesses of formal theories

In this section, I have provided an overview of formal theories that have made use of different sets of tools to model people's responsibility attributions. All theories acknowledge that attributions of responsibility are closely linked to causality. I have argued that formal theories of responsibility attribution should hence be built upon a theory that can adequately represent people's causal understanding of the world. On these grounds, I have dismissed covariational, connectionist and logic-based accounts of responsibility attribution.

Counterfactual theories, in contrast, do take people's causal understanding of the

[22]See Gopnik and Wellman (in press) and Schulz (2012) for comprehensive reviews of the developmental literature on causal learning and inference.

world into account. They argue that an event will be held responsible to the extent that it was perceived to have made a difference to the outcome. They express this 'difference' in terms of comparing the probability of the outcome in the presence of the cause with the counterfactual probability that this outcome would have prevailed in the absence of the cause (Brewer, 1977; Spellman, 1997). Fincham and Jaspars (1983) have added to this basic notion of making a difference that it also matters whether the observer thinks that other people would have acted in the same way. Finally, Petrocelli et al. (2011) have argued that attributions of responsibility are not only sensitive to whether a relevant counterfactual would have undone the outcome or not but also to what the perceived chances were that such a counterfactual state of affairs would have come about in the first place.

However, as I have argued, the formal accounts discussed thus far have not made the connection between causality and counterfactuals explicit. In their experiments, researchers have relied on participants generating the appropriate counterfactuals but they have not provided an account of how people use their causal knowledge to reason counterfactually. The framework of causal Bayesian networks provides such a formalism and I will show in the next chapter how it can be used to derive a formal model of responsibility attribution that is grounded in people's causal representation of the world.

2.4 Conclusion

Let me conclude this section by comparing the strengths and weaknesses of label theories and formal theories side by side (see Table 2.1). Label theories acknowledge the richness of concepts that are involved in the attribution process (see Figures 2.3, 2.5 and 2.6). Attributions of responsibility are dependent upon a diversity of factors such as the actor's perceived control, the foreseeability of the outcome and whether or not the outcome was intended. Both normative and descriptive label theories have been developed. As we have seen, these theories differ in the way in which attributions of causality, responsibility and blame/praise are related. Whereas normative theories adhere to the notion of entailment (i.e. causality before responsibility and blame), descriptive theories highlight that judgments of causality can be influenced by moral considerations. Label theories suffer from a lack of precision which can make empirical falsification difficult and thus potentially hamper progress in the field. The fact that predictions derived from label theories tend to be qualitative rather than quantitative, is reflected in the methodological approach which mostly relies on vignette studies in which qualitative aspects of the situation are varied (e.g. whether or not the actor foresaw the outcome).

Formal theories, in contrast, provide precise definitions of the factors that are assumed to influence attributions of responsibility. Furthermore, these theories specify the ways in which the different factors combine to determine the attribution (e.g. in an additive or multiplicative manner). The theories generate quantitative predictions

that can be tested in richer experimental setups that go beyond the binary distinction of whether a factor is absent or present (see, e.g. Spellman, 1997). Formal models can also be used to explore sources for individual differences between participants. A counterfactual model, for example, predicts that two people will reach a different causal judgment when their subjective beliefs about what would have happened in the relevant counterfactual world are different.

Due to the precise formulation of these theories, they facilitate empirical falsification that leads to theoretical refinements which help to move the field forward (Mandel, 2003; McClure et al., 2007). However, the challenge of capturing the richness of factors that influence attributions remains. For example, up until now, there is no formal model which specifies in a quantitative way how intentions affect attributions of responsibility. It is worth nothing that quantitative precision can also be a danger, especially if the formalisation is inadequate. Much of the early research in attribution theory has been dominated by Kelley's (1973) covariational approach despite the fact that covariation is at best a cue to causality but not to be confused with causality itself (cf. Kelley, 1983). That is, there is a danger to see the world through the glasses of one's own favourite formal framework.

Despite the fact that both label theories and formal theories acknowledge the importance of causality for attributions of responsibility, neither has provided an adequate account of how people represent the causal structure of the world. Being precise about

Table 2.1: Strengths and weaknesses of label theories versus formal theories of responsibility attribution; ✓= strength, ✗= weakness.

	Label theories	Formal theories
conceptual richness	✓	✗
distinguish causality, responsibility and blame	✓	✗
direct role of intentions	✓	✗
normative / descriptive distinction	✓	✗
high precision of concepts	✗	✓
quantitative predictions	✗	✓
possible to model individual differences	✗	✓
empirical falsifiability	✗	✓
based on causal model of the situation	✗	✗
individual responsibility in groups	✗	✗
methodological diversity	✗	✗

the causal structure of the situation becomes especially important in situations in which there are multiple causes (or individuals) that jointly contribute to an outcome (Kelley, 1983). None of the theories discussed thus far have made predictions about to what degree individuals should be held responsible for an outcome that has been brought about collectively. In such situations, outcomes are often causally overdetermined in that the same outcome would have prevailed even if one of the causes had been absent. These situations are difficult to handle for simple counterfactual models. Chapter 3 develops a formal account that deals with such cases. Chapter 4 explores the predictions of this account by going beyond the usual experimental technique of having participants attribute responsibility to protagonists in vignettes.

To sum up, in the current landscape of models on attribution, there appears to be a trade-off between capturing the richness of causal factors involved on the one hand and making precise how different factors are expected to influence attributions on the other hand. While some have started to develop richer formal models of attribution (Baker et al., 2009, 2011; Goodman et al., 2009; Sloman, Barbey, & Hotaling, 2009; Wolff, 2003; Wolff, Barbey, & Hausknecht, 2010) much more work remains to be done ...

Chapter 3

Causality, Counterfactuals and Responsibility

> We have come to think of the actual as one among many possible worlds. We need to repaint that picture. All possible worlds lie within the actual one.
>
> – **Goodman (1983, p. 57)**

In this chapter, I will lay the theoretical groundwork on which subsequent empirical chapters build. I will argue for a close relationship between causality, counterfactuals and attributions of responsibility. All the theories of responsibility attribution that I have introduced in the preceding chapter, have highlighted the role of causality for attributions of responsibility. Some of the theories have also argued for a close correspondence between considerations about counterfactuals and responsibility attributions (Brewer, 1977; Petrocelli et al., 2011; Spellman, 1997). These accounts share the intuition that attributions of responsibility are related to the degree to which the event under consideration made a difference to the outcome.

We can think not only about how our actions will make a difference in the future but also about how events in the past shaped the situation we find ourselves in at present. I might wonder, for example, what would have happened if I had decided not to do a PhD. Presumably, I would have found myself in a very different world from the one that I am currently experiencing. This example leaves room for a vast space of possible worlds which is in part due to the temporal distance of the present (i.e. sitting in front of my computer and trying to write a catchy introduction to this chapter) from the considered time point in which I would have entered a different counterfactual world (i.e. contemplating what to do with my life on a trip in India before having accepted a PhD position at UCL). However, in many situations the relevant space of counterfactual

worlds to consider is much smaller. Furthermore, we are often interested in evaluating the chances that a particular change in the past (e.g. taking the bike to work instead of the tube) would have led to a particular change in the present (e.g. making it to a meeting in time rather than being late).

While the models outlined in the previous chapter acknowledge the relationship between counterfactuals and attributions of responsibility, they have not succeeded in making the nature of this relationship sufficiently precise. For example, Spellman (1997) argues that attributions of responsibility are related to the degree to which an individual contribution changed the probability of the eventual outcome. In all her experiments, participants were provided with all the information they needed in order to do this calculation, that is, they were given the probabilities of the effect for each possible combination of causal events. However, in real life, we do not have access to these probabilities. Hence, we need a principled way of assessing the likely consequences that changes of the past would have resulted in. In this chapter, I will argue that this is impossible unless we take causality seriously.

Philosophical theories of causation can be broadly divided into *reductionist* and *non-reductionist* theories. According to reductionists, causation can be reduced to another more fundamental notion such as counterfactual or probabilistic dependence (e.g. Suppes, 1970), regularity (e.g. Hume, 1748/1975) or the transmission of physical quantities (e.g. Wolff, 2007). Non-reductionists, in contrast, take the notion of causality as fundamental and do not aim to reduce it to something else (see Hitchcock, 2001a; Pearl, 2000; Woodward, 2003). Rather, these theorists define causality in terms of (counterfactual) interventions and propose that what it means for A to be a cause of B is that there are potential interventions on A which would change the value (or probability distribution) of B. Interventionists theories of causation are non-reductionist because they rely on an idealised notion of intervention (as a local change in a system that leaves all other aspects of the system intact) that is itself causal.

While this might look hopelessly circular on first sight (defining causation in terms of a causal notion of intervention), we will see below that such a theoretical approach is very powerful as a psychological theory of causation. It avoids problems that counterfactual theories face which aim to ground counterfactuals without appeal to causal knowledge (see Edgington, 2011; Woodward, 2011b). Indeed, the interventionist conception of causality supports a rigorous analysis of other concepts related to causation such as prediction, explanation, counterfactual reasoning as well as judgments about actual causation and responsibility (see Schaffer, 2003).

Recently, interventionist theories of causation have gained popularity due to the theoretical work of computer scientists who developed a causal modelling framework that clarifies the relationships between causal models, probability, interventions and counterfactuals (Pearl, 1988, 2000; Sloman, 2005; Spirtes et al., 2000). This work has had an enormous influence on the field of cognitive science. Here, I will argue (and in

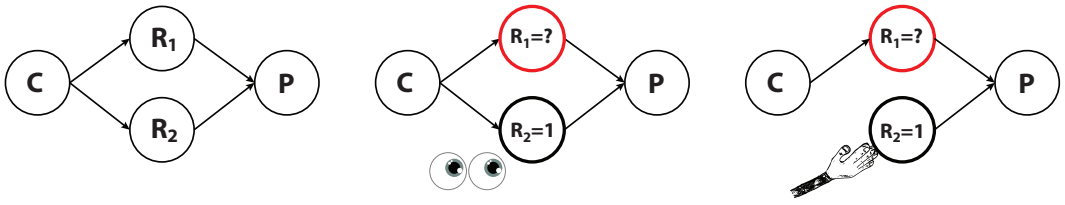
Chapter 4, I will show) that this work also has much to say about how people attribute responsibility.

The rest of the chapter is organised in the following way. I will first outline Pearl's (2000) framework of representing causality in terms of causal Bayes nets. I will then illustrate how causal Bayes nets can be used to answer counterfactual queries and to reason about cases of actual causation. Finally, I will show how we can formally express queries about responsibility in this framework.

3.1 Causal Bayes Nets

In the interventionist approach to causation, causal claims are embedded in networks of causal relationships (Woodward, 2003). The causal relata in such an approach are variables, whereby each variable represents a set of different possible events. A causal Bayes net (CBN) consists of a graphical model and a set of structural equations which describes the relationships between the variables in the model. The variables are depicted as nodes in the network and the causal dependencies between variables are indicated by arrows (see Figure 3.1a). Traditionally, CBNs are restricted to the class of directed acyclic graphs in which the causal relationship are only one-directional, that is, feedback loops are not allowed in the network. In order to simplify the discussion, we will focus on CBNs with deterministic, generative relationships.[1] In my discussion, I will use a simplified example from Pearl (2000, Chapter 7) which fits well with the theme of blame attribution.

Consider the following situation: A commander C is in charge of a small firing squad consisting of rifleman R_1 and rifleman R_2 whose job it is to execute a prisoner P. We assume for simplicity that we can represent each causal actor in our scenario with binary variables. Hence, C can either give the order to *shoot* ($C = 1$) or *not to shoot* ($C = 0$). Each of the riflemen R_1 and R_2 can either *shoot* ($R = 1$) or *not shoot* ($R = 0$). Finally, the prisoner, the poor chap, can either *die* ($P = 1$) or *survive* ($P = 0$).



(a) Causal Network **(b)** Observation **(c)** Intervention

Figure 3.1: (a) Causal network representation of the firing squad scenario, (b) Observing an event, (c) Intervening in the network to bring about an event.

[1] See Pearl (2000) for a comprehensive discussion of CBNs with probabilistic relationships that can include preventive dependencies between variables.

Figure 3.1a shows a causal network representation of the firing squad scenario. From the arrows in the network we can already read off some of the assumptions that this representation makes about the causal relationships between the variables. Both of the riflemen are causally dependent upon the commander and the prisoner's fate is causally dependent upon the riflemen. Furthermore, the fact that there is no directed arrow between the two riflemen shows that they do not influence each other directly. The fact that there is no direct arrow from the commander to the prisoner shows that the commander cannot bring about the death of the prisoner directly but only through the actions of the riflemen (i.e. blowing the whistle won't kill the prisoner).

While knowing about the arrows in the network is sufficient to read off the general causal relationships between the variables, the graph structure obscures the exact ways in which the different variables depend upon each other. For example, we see in the graph that both riflemen exert a causal influence on the prisoner. However, we cannot infer directly from the graph how the causal influences of both riflemen combine. Hence, we need to inspect the structural equations that describe the local relationships between the variables. For each variable in the model, there is a structural equation which determines its value by the status of its *parent* variables and the value of exogenous factors that are not explicitly represented in the model.[2] The assumption that we can express the joint probability distribution over all variables in the network in terms of local conditional distributions over children given their parents is a core assumption in a CBN called the *Markov condition*. Thus, in order to infer the value of P in our network, we only need to consider the values of P's parents R_1 and R_2 and not the value of C. In other words, our model expresses the assumption that P is independent of C given the values of R_1 and R_2.

Imagine that the prisoner is hidden behind a wall and we are wondering whether he is dead or not. Intuitively, knowing whether or not the two riflemen shot entails all the information we need in order to judge whether the prisoner died. Learning that the commander gave an order to shoot, given that we already know what the riflemen did, does not change our belief about whether or not the prisoner died. In CBN terminology, our knowledge about the values of R_1 and R_2 screens off P from C – the information flow (or dependence) from P to C is blocked by knowing about R_1 and R_2.

In sum, a CBN represents the information about the joint probability distribution over the set possible worlds that are licensed by the model in terms of autonomous causal mechanisms that hold between *parents* and *children*. Conversely, knowledge about these local mechanisms between variables entails all the information we need in order to generate the complete joint probability distribution (in CBN terminology: the joint probability distribution *factorises* into a set of conditional probability distributions).

For our particular representation, we will assume that there are no external factors

[2]Kinship terms are used to describe the causal relationships between variables in the network. For example, in our network P has two *parents* R_1 and R_2 which are in turn the *children* of C.

that can influence the values of the non-root variables in the network.[3] Equation 3.1 shows the set of structural equations that describe the causal mechanisms underlying the causal network in Figure 3.1a. The equal signs in the structural equations are not to be confused with mathematical equality but should be regarded as assignments of values to variables similar to standard practice in computer programming. In other words, the variables on the left hand side are the effects (children in the graph) and the variables on the right hand side are the causes (parents in the graph) of a local causal mechanism.

$$\begin{aligned} & & (C) \\ R_1 &= C & (R_1) \\ R_2 &= C & (R_2) \\ P &= max(R_1, R_2) & (P) \end{aligned} \tag{3.1}$$

We can see that values of the two riflemen are determined by the value of the commander. That is, whenever the commander says shoot ($C = 1$) each of the riflemen shoots ($R_1 = 1$ and $R_2 = 1$). The structural equation for P, which is a common effect of R_1 and R_2 shows that the causal influences of both of the riflemen combine in a disjunctive fashion. Hence, P's value is 1 if either $R_1 = 1$ or $R_2 = 1$ or both. Finally, we can also see from the structural equations that the value of C is determined by factors that are external to variables specified in the model.

The causal network together with the structural equations captures the causal implications of the firing squad scenario. While the model is constructed in order to represent the specific firing squad scenario described above, it exhibits some degree of generality. The structural equations constrain the values that the variables can take on. However, we have not determined in advance which values the variables have actually realised. For our representation of the firing squad scenario, there are two possible sets of assignments of values to variables. Either the values of all variables are 1 (the commander gives the order to shoot, both riflemen shoot and the prisoner dies) or 0 (the commander gives the order not to shoot, both riflemen do not shoot and the prisoner survives). Thus, a CBN captures the causal structure of the world and hence implicitly represents a set (or distribution) of worlds that are consistent with the causal structure.

A CBN supports both *type* and *token* causal claims (e.g Hall, 2007; Hitchcock, 2001a, 2009). Claims about *type causation* concern general relationships between types of events such as *gun shots cause deaths*. Applied to our example, the model incorporates the type causal assumptions that commanders generally cause riflemen to shoot and that prisoners die when riflemen shoot. Type causation is related to prediction and intervention. What would happen if one of the riflemen were to shoot? What would

[3]Root variables are variables that do not have any parents, that is, variables that do not have any incoming links. In our causal network, C is the only root variable.

happen if we issued a command to shoot (e.g. by making a sound that sounds just like the commander's whistle)?

Claims about *token causation*, in contrast, concern statements about the actual causes of an outcome in a particular situation such as *Jim's gun shot caused Joe's death*. Once a particular world has been instantiated, we can ask who actually caused the death of the prisoner. Consider that the commander ordered the riflemen to shoot, both riflemen shot and the prisoner died. Did the commander actually cause the prisoner's death? Did each of the riflemen cause the death? Claims about token causation are intimately linked with diagnostic inferences, that is, inferences about unknown causes from known effects. Similarly, token causation is closely linked to counterfactual claims. We might ask, for example, whether the prisoner *would* have died if one of the riflemen had not shot.

Thus, we see that a CBN captures how information flows in two directions: *forward* from causes to effects and *backward* from effects to causes. We will see in Chapter 4 how both the *forward* and *backward* looking nature of causality are crucial for people's judgments of responsibility. Now that we have specified the exact causal relationships between the variables which incorporate our assumptions about the firing squad scenario, we can use this representation for different types of inference: inference based on observation, intervention and counterfactuals.

3.1.1 Observation, intervention & counterfactuals

One of the main reasons for the popularity of CBNs in artificial intelligence and psychology is their ability to differentiate between inferences based on observation versus intervention (Sloman & Lagnado, 2005). Imagine that you do not have perceptual access to all of the actors in the firing squad scenario but you only observed that the second rifleman shot (see Figure 3.1b). Given this piece of information what can you infer about the values of the other variables in the causal network? We assume that you know about the exact ways in which the different causal variables are connected (i.e. you know everything that is specified in Equation 3.1). Given that you have observed that rifleman 2 has shot, what do you think are the chances that rifleman 1 shot as well $P(R_1|R_2)$? As can be seen in the network, we cannot answer this question directly as there is no direct relationship between the two riflemen. However, the two riflemen are connected in the network through two different routes, namely through $R_1 \leftarrow C \rightarrow R_2$ and through $R_1 \rightarrow P \leftarrow R_2$. As we have seen, a CBN supports reasoning in two directions: predictive reasoning from causes to effects (or from parents to children in a graph) and diagnostic reasoning from effects to causes (or from children to parents). Hence, given that we observed $R_2 = 1$ we can reason diagnostically along the first route to infer that C must be 1. As Equation 3.1 shows $R_2 = 1$ if and only if $C = 1$. This follows given our assumption that riflemen only shoot when the commander says so. From the

fact that $C = 1$ we can then infer that R_1 must also be 1.

Let us contrast the situation in which we have observed that rifleman 2 shot with a situation in which we actively intervene in the system. With the help of some of UCL's Institute of Cognitive Neuroscience researchers, we secretly implanted a chip into rifleman 2's brain that, if turned on, causes him to shoot. We turn on the chip and it causes a current in rifleman 2's brain which makes him shoot. What can we infer now about whether rifleman 1 shot? Intuitively, we are now more uncertain about whether he shot. In order to capture the difference between observation and intervention, Pearl (2000) introduced what he calls the *do*-operator. Here we are interested in the probability that rifleman 1 shot given that we intervened in the system to make rifleman 2 shoot $P(R_1|do\ R_2)$. According to the *do*-operator, we intervene in a causal system by locally breaking the causal mechanisms that normally determine the way in which the system operates. More specifically, an intervention is represented as an operation that is external to the system and fixes the intervened-on variable to a value independently of the values of the other variables in the network.

Graphically, the consequences of an intervention can be represented by constructing a new causal network through a process called graph-surgery: the original graph (see Figure 3.1a) is mutilated to generate a new graph (see Figure 3.1c) in which all incoming links to the intervened-on variable are removed. All the other relationships remain intact. In terms of the structural equations that are associated with the new graph, we replace R_2 in all equations on the left hand side with the value to which we have set the variable (see Equation 3.2).

Given our new graph and the new set of equations, we can now apply the same rationale as above. We know that $R_2 = 1$ due to our intervention. However, since R_2 is now independent of C, we cannot infer anything about C from the value of R_2. The probability that C is positive is given by the circumstances which are external to our model representation. It follows, that we can also not be sure whether rifleman 1 shot. This depends on whether or not the commander gave the order to shoot at the same time as we activated the chip in rifleman 2's brain.

$$
\begin{aligned}
& && (C) \\
& R_1 = C && (R_1) \\
& R_2 = 1 && (R_2) \\
& P = max(R_1, R_2) && (P)
\end{aligned} \tag{3.2}
$$

Hence, we see that observations and interventions license different inferences about the states of the other variables in the system. Whereas we can reason backwards from effects to causes when we have observed these effects, this inference is not valid when we

have actively intervened in the system. The value of a variable which was set through intervention is not diagnostic anymore of the value of its parents. Sloman and Lagnado (2005) have shown that people are sensitive to the different implications of observations and interventions for the status of the other variables in the system.[4]

As outlined above, one of the features of CBNs is that they also support counterfactual reasoning. That is, reasoning about how the present could have turned out differently, had an event in the past been different. For example, given that we observed the prisoner's death $P = 1$, we might wonder whether the prisoner would still have died if rifleman 1 had refused the order and decided not to shoot. That is, we want to assess the truth of the counterfactual statement: *If rifleman 1 had not shot, then the prisoner would have survived.* In order to evaluate counterfactual statements using a CBN, we have to combine what we have learned about observation and intervention.[5] According to Pearl (2000), the evaluation of counterfactuals involves three steps: 1) abduction, 2) intervention and 3) prediction (see Figure 3.2).

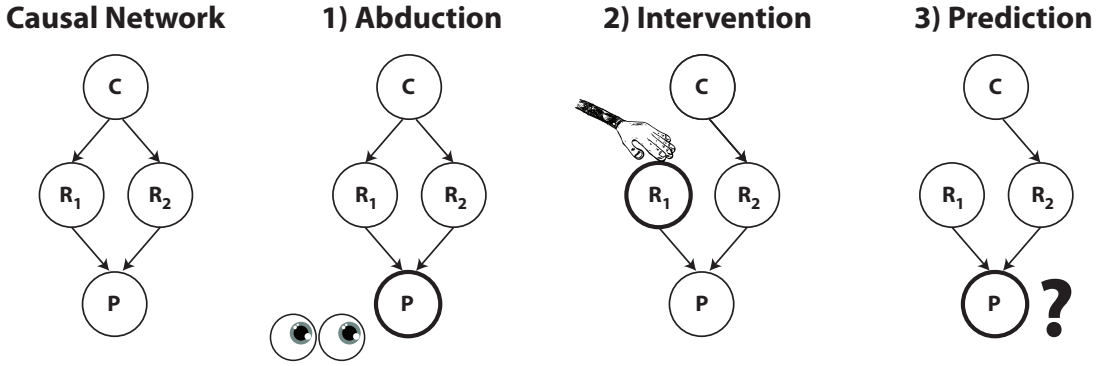


Figure 3.2: Evaluation of counterfactuals according to Pearl (2000).

In the *abduction step*, we fix the values of the variables so that they are in line with our observations. From the fact that the prisoner is dead and our knowledge about the deterministic relationship between the variables as specified in Equation 3.1 we can infer that the value of each variable must have been 1. This includes the only random observation in our network, namely the positive value of C which is determined through factors outside the causal network.

In the *intervention step*, we generate a twin model that is identical to the original model (i.e. including all the random factors) but in which we intervene on the target variable in the antecedent of our counterfactual (see Equation 3.3). We use the *do*-operator to fix the target variable which creates a novel counterfactual model. Finally, in the *prediction step*, we then reevaluate this novel model to infer the values of the

[4]However, see Cartwright (1995) for a criticism of the modularity assumption which states that a system can be composed of local autonomous mechanisms and that interventions in the system can be performed without disturbing the rest of the causal system. Also, see Pearl (2008) for a response.

[5]Here I will focus on the deterministic case. The probabilistic case will be discussed in Chapter 6.

other variables in the network (given our counterfactual intervention).

$$\begin{aligned} C & \qquad (C) \\ R_1 = 0 & \qquad (R_1) \\ R_2 = C & \qquad (R_2) \\ P = max(R_1, R_2) & \qquad (P) \end{aligned} \tag{3.3}$$

Through the counterfactual intervention, we have broken the link between C and R_1. However, our local counterfactual intervention has left the rest of the network untouched. Given that we copied all the random events that were observed (and inferred) in the original situation, we know that $C = 1$ and $R_2 = 1$. The only variable we need to reevaluate in the prediction step is P. Since P is 1 if either $R_1 = 1$ or $R_2 = 1$ or both, we know that the counterfactual statement is false. That is, the prisoner would have died even if rifleman 1 had refused to shoot because rifleman 2 would still have shot. In contrast, the following counterfactual would turn out to be true according to our model: *If the commander had not given the commando to shoot, the prisoner would not have died.*

Let me highlight the difference between counterfactuals and interventions once more. When evaluating a counterfactual statement, we know what values all (or some) of the variables had in the actual situation. We then go back and consider what would have happened, if a certain variable in the network would have taken a different value. According to Pearl (2000) only variables that are downstream from the variable for which we consider an alternative value are potentially affected by this intervention. The other variables are left untouched at their original values.[6] In contrast, when we intervene in a system, we don't know the values of the downstream variables in the network yet. Hence, considering an intervention which prevents R_1 from shooting is different from considering what would have happened if R_1 would not have shot. In the intervention case, we cannot be sure whether P *will die* or not – it depends on the status of R_2. However, in the counterfactual case, we can be sure that P *would have died* anyhow, from knowing that R_1 shot we can infer that R_2 must have shot as well.

We have seen how a single network representation supports inference based on different types of evidence: passively observing the system, actively intervening on the system and considering counterfactual events. In the following section, we will turn to the relationship between counterfactuals and attributions of responsibility.

[6]See Hiddleston (2005) for an alternative approach of modelling counterfactual conditionals and Rips (2010) for an empirical test of the two competing accounts.

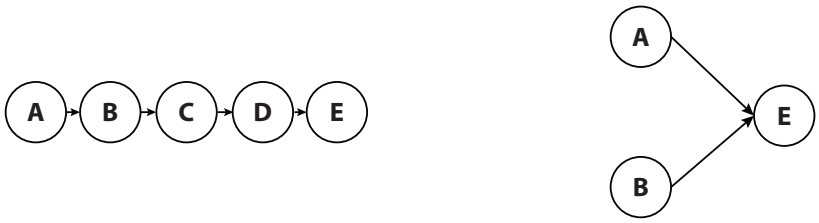
3.2 From Counterfactuals to Attributions of Responsibility

As outlined in the previous chapter, many have argued that responsibility attributions are closely related to counterfactual considerations (Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Petrocelli et al., 2011; Spellman & Kincannon, 2001; Wells & Gavanski, 1989; Wells, Taylor, & Turtle, 1987). In law, one of the conditions for an event to qualify as a cause-in-fact is the *sine qua non* condition (Spellman & Kincannon, 2001). According to this condition, an event only qualifies as a cause if the effect would not have happened *but for* the event under consideration. Indeed, the intuition is strong that an actor will not be held responsible for an outcome if the outcome would have happened irrespective of what the person did.

While there appears to be a strong relationship between counterfactual dependence and whether a person should be held responsible, it is easy to see that the strict notion of counterfactual dependence fails as a criterion for whether an event qualifies as being responsible for the outcome. Indeed, the *but for* test is both too weak (i.e. inclusive) and too strong (i.e. exclusive): (i) too weak because it rules in many events as causes that we would not want to hold responsible and (ii) too strong because it excludes causes that we consider responsible.

The *but for* test is too weak because very (causally) remote events would count as causes for effects. However, the intuition is that in order to be held responsible, the event under consideration has to be sufficiently proximal to the effect of interest. For example, consider the causal chain situation depicted in Figure 3.3a. An innocent person E was mistakenly killed by an assassin D who received his orders from C. B had bought the weapon from A and passed it on to C. In this situation, would we judge A to be responsible for E's death? The intuitions is strong that the answer is no (cf. Halpern & Hitchcock, forthcoming). While A, as a seller of weapons, could in general be held responsible for people dying (a *type* causal claim), his action seems to be too remote to qualify as the cause of E's death (a *token* causal claim). Indeed, it would be easy to generate situations in which the connection was even more remote (i.e. consider the worker in the steel manufacture who built a part that was used for the gun with which D shot E). Hence, it seems that in many situations the degree of responsibility reduces with an increased distance to the eventual effect.

Some research has shown that when later variables in a causal system strongly depend upon earlier variables then people have a tendency to attribute more responsibility to the first event that initiated the chain (Wells et al., 1987). In contrast, if the relationships between the variables are rather loose, people tend to attribute the most responsibility to the final event (Miller & Gunasegaram, 1990). For example, if A is a Mafia boss who wants to see E dead and $B - D$ are his subordinates who are more or less coerced to follow his orders then A would be seen as responsible for E's death. However, if A



(a) Causal Chain **(b)** Common Effect

Figure 3.3: Different causal structures that imply different degrees of responsibility for A.

is a traffic warden who signals B to wait in his car for a while and B then later gets the last free parking space just before C who in turn needs to continue driving around. Because of his slow driving, C causes D to do a lane change who then happens to run over a dog E. In this situation it seems clear that A does not have any responsibility for E's death, despite the fact that E would not have died (at least not in this way and at that time) but for A's action.

However, as mentioned above, the *but for* test is not only too weak a criterion for whether an event can be held responsible or not but it is also too strong at the same time. Some events that we feel deserve responsibility for an outcome are ruled out. Consider the network structure in Figure 3.3b. A and B are assassins who simultaneously shoot at the victim E who dies as a result. As it turns out, E's death was overdetermined. Each assassin's shot would have individually been sufficient to cause the victim's death.[7] If we now use the *but for* test as a guide for whether A is responsible for E's death we see that it fails. E would have died *but for* A's action (because of B). The same is true for B of course. Hence, we have a dead victim without anyone being responsible for it – something is clearly wrong.

3.2.1 A structural model of responsibility attribution

In the previous section, we have seen that the counterfactual *but for* test is both too weak and too strong as a criterion for assigning responsibility. The structural model of responsibility attribution (Chockler & Halpern, 2004) addresses both problems. First, it relaxes the strict notion of counterfactual dependence. Second, it incorporates the extent to which an agent could foresee the consequences of their action. The first fix makes the model more inclusive compared to the *but for* test: events that are excluded as causes by the *but for* test can count as causes in the structural model and receive (partial) responsibility. Including the notion of foresight makes the model more exclusive: events

[7] For a real-world case in which several policemen killed a suspect in an overdetermined fashion in London, 6th May 2008, see: `http://www.ipcc.gov.uk/documents/investigation_commissioner_reports/saunders-21-12-10.pdf`

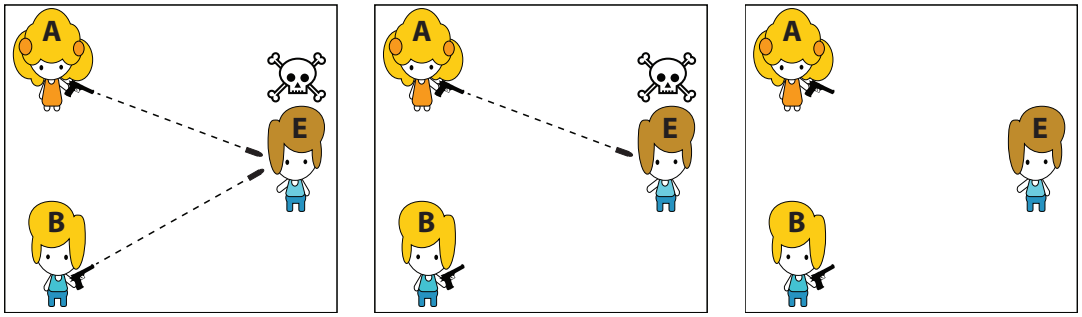
that would count as causes according to the *but for* test will only receive minimal blame.

3.2.1.1 Ruling in *responsible* causes

Let us first consider the way in which the structural model rules in events as causes that the *but for* test misses. Figure 3.4 shows three of the four possible situations which could come about for the overdetermination scenario described above. We have seen that according to the simple counterfactual criterion, neither of the events A or B count as a cause for E in the situation in which either A or B would have individually been sufficient to bring about E (see Figure 3.4a).

In order to solve this problem, the structural model employs a relaxed criterion of counterfactual dependence (see Halpern & Hitchcock, 2011; Halpern & Pearl, 2005) according to which an event A can still count as an actual cause of another event E even when there was no counterfactual dependence (in the *but for* sense) between the two events in the actual situation. In order for an event to count as an actual cause, there merely has to be a *possible situation* in which the *but for* test would have held (see also Hall, 2007; Hitchcock, 2001a; Woodward, 2003; Yablo, 2002, for similar ideas of rescuing counterfactual accounts of actual causation). Thus, actual causation is defined in terms of counterfactual dependence *under certain contingencies*.

For example, assume that the effect event E was causally overdetermined by the presence of A and B. While there is no counterfactual dependence between A and E in the actual situation, there is a possible situation in which A would have been a *but for* cause of E. Namely, the situation in which event B would not have occurred (see Figure 3.4b). In this possible situation, it is true that E would not have occurred but for A (see Figure 3.4c).



(a) Both shoot **(b)** Only A shoots **(c)** Neither shoots

Figure 3.4: Three possible worlds that could have come about in the overdetermination scenario.

Chockler and Halpern (2004) impose some constraints that limit the possible contingencies that can be considered in order to check whether an event under consideration should count as an actual cause. One such constraint states, roughly, that fixing the

value of an off-path variable B, in order to assess whether C counts as an actual cause of E should not affect any of the variables on the causal path from C to E. In other words, we are only allowed to consider contingencies in which holding certain variables fixed does not interfere with the actual causal path of interest. However, for the cases with which we are going to be concerned in this chapter and in Chapter 4 we need not worry about these subtleties. Here, we will only deal with situations in which multiple causes bring about an effect in an independent fashion. We will return to this point in Chapter 6 which deals with actual causation in a richer context.

In order to facilitate the discussion of some of the more complex situations, it is helpful to introduce the notion of *pivotality*. An event (or agent) is pivotal in a situation if there is a counterfactual dependence between the effect event of interest and the cause event (or causal agent) under consideration. That is, if the *but for* test is positive for an event in a given situation, this event was pivotal for the outcome. The examples we use in the following will employ agents who can either succeed or fail in their individual tasks and who collectively either bring about a positive or negative group outcome.[8]

From the relaxed notion of counterfactual dependence that allows agents to have caused the outcome even if they were not pivotal in the actual situation, Chockler and Halpern (2004) derive their notion of responsibility. The core idea is that responsibility does not drop to zero as soon as an agent is not pivotal anymore (as would be suggested by the strict *but for* test) but rather reduces gradually the more an outcome is overdetermined. An agent is fully responsible if she is pivotal. Responsibility then decreases with the causal (dis-)similarity of the actual situation to a possible situation in which the event of consideration would have been pivotal. The causal similarity between two situations is given by the minimal number of variables whose value needs to be changed in order to transform one situation into the other. The closer a possible situation in which the agent would have been pivotal is to the actual situation, the more responsible the agent is for the outcome.

The formal definition of how much responsibility a cause event c, as an instantiation of the cause variable C, has for a particular instantiation e of an effect variable E is given by

$$\mathit{Responsibility}(C = c, E = e) = \frac{1}{N+1}, \tag{3.4}$$

where N equals the minimal number of changes that have to be made to transform the actual situation into a possible situation in which the values of E and C would have been counterfactually dependent.

[8]It is worth pointing out that the structural model of responsibility applies equally well to physical objects such as parts of a mechanical system (Chockler, Halpern, & Kupferman, 2008). For example, a faulty cylinder might be responsible for a broken engine.

If an agent is pivotal in the actual situation, the number of changes N equals zero and the agent is fully responsible. The larger N, the less responsible c is for e (see Figure 3.5). For example, if the values of two variables would need to be changed to transform the actual situation into a possible situation in which the agent of interest would have been pivotal, the agent's responsibility for the outcome would be $\frac{1}{3}$. Note that according to the structural model, responsibility does not decrease linearly but logistically as a function of the number of changes N. Hence, the decrease of responsibility is approximately exponential for small number of changes and asymptotes to zero for larger number of changes. For example, moving from zero changes to one change, results in a responsibility reduction from 1 to $\frac{1}{2}$. In contrast, moving from three changes to four changes only results in a reduction of responsibility from $\frac{1}{4}$ to $\frac{1}{5}$.

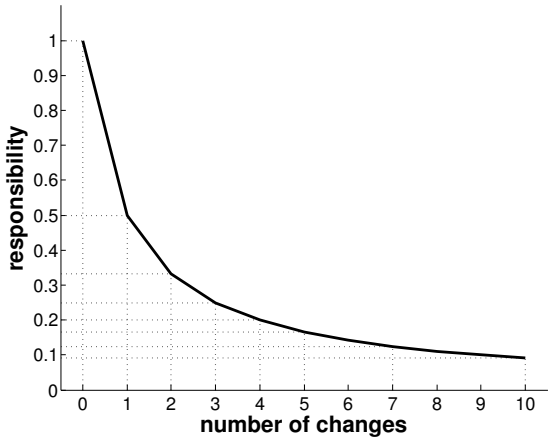
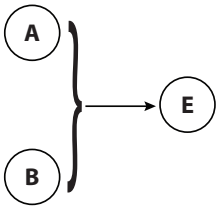


Figure 3.5: Responsibility function: $Responsibility(c, e) = \frac{1}{N+1}$.

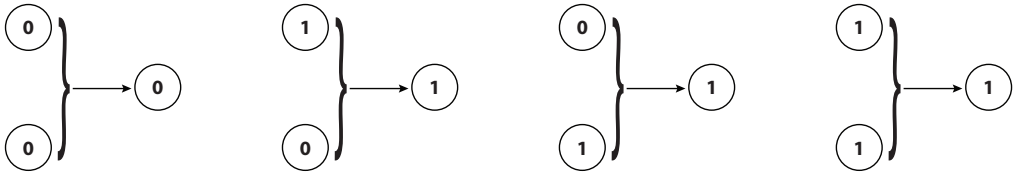
Two simple examples of how the model works In this section, I will illustrate some examples of how the structural model of responsibility attribution works (Chockler & Halpern, 2004). We will start with fairly simple examples before moving on to some more complex ones.

Let us start with the situation we have already described above in which there are two causes, A and B that together contribute to an effect E. The nature of the effect is such that it is positive if at least one of the causes is positive (i.e. $E = max(A, B)$). To make things a bit more vivid, let us imagine that A and B are players in a little game who contribute to their team outcome E. Further, for simplicity, we will assume that all variables are binary. A and B can either *succeed* or *fail* in their individual tasks and, depending on how they perform, their team either *wins* or *loses*. Figure 3.6 gives a graphical representation of the causal structure with the four possible situations that can occur. In our graphs, we will use a curly bracket to indicate that the performances of two players combine in a disjunctive fashion.

Let us focus on player A and see how much responsibility he receives according to



(a) Disjunctive Structure



(b) Both fail, $Resp(A = 0, E = 0) = 1$

(c) A succeeds, $Resp(A = 1, E = 1) = 1$

(d) B succeeds, $Resp(A = 0, E = 1) = 0$

(e) Both succeed, $Resp(A = 1, E = 1) = \frac{1}{2}$

Figure 3.6: Disjunctive causal structure (top) with the four possible situations that can occur (bottom). *Note*: The curly bracket indicates that only at least one of the causes (A or B) needs to be positive in order to bring about the effect (E).

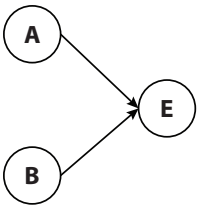
the structural model in the four different situations. In the situation in which both fail and the team lost (i.e. $E = 0$, see Figure 3.6b) A has a responsibility of 1 since A is pivotal in the actual situation. The team outcome E would have been different had A succeeded rather than failed. Similarly, in the situation in which only A succeeded and the team won (i.e. $E = 1$, see Figure 3.6c), A has a responsibility of 1. The team would have lost, had A failed in his task. In the situation in which only B succeeded (see Figure 3.6d), A has a responsibility of 0. A does not count as a cause of the outcome in this situation. There is no counterfactual dependence between E and A and if we were to change the value of B in order to create a counterfactual dependence, the value of E would change from $E = 1$ to $E = 0$. However, in assessing A's responsibility, changes to other variables which alter the value of the effect are invalid.[9] Given the structural description of the situation, A cannot make the team *lose* through *succeeding* in his task. As a rule of thumb, for the situations we will look at, A's responsibility is 0 if the value of A and E mismatch.[10] Finally, let us consider the situation in which both

[9]This is another constraint imposed on possible contingencies (see Halpern & Pearl, 2005, p. 9). According to the structural model, responsibility is always outcome dependent. That is, a cause event is responsible for the particular value of an effect event. Expressed differently, we want to assess whether changing A from its actual a_{actual} value to another possible value $a_{possible}$ could change E from its actual value e_{actual} to another possible value $e_{possible}$. Counterfactual contingencies in which the effect event would have been different are excluded from the set of possible situations in which the pivotality of the cause event of interest is assessed. In situations in which there is no possible contingency that would render the two variables of interest counterfactually dependent, the number of changes is fixed to infinity ∞ and hence the degree of responsibility reduces to 0.

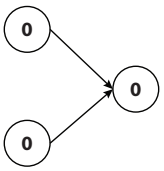
[10]If, in contrast, the individual performances of A and B had combined according to a exclusive-or (XOR) function (i.e. $E = (A \wedge \neg B) \vee (\neg A \wedge B)$), then A's responsibility would have been 1. In fact, in

succeeded in their individual tasks and the team won (see Figure 3.6e). While A is not pivotal in the actual situation, he would have been pivotal in the possible situation in which B had failed (cf. Figure 3.6c). Since one variable needs to be changed ($N = 1$) in order to transform the actual situation into this possible situation, A's responsibility equals $\frac{1}{1+1} = \frac{1}{2}$.

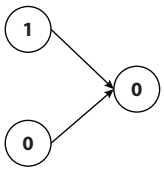
Let us now consider a different situation in which the contributions of two players combine in a *conjunctive* fashion (see Figure 3.7). That is, in order for the team to win, both players have to succeed in their individual tasks ($E = min(A, B)$). Graphically, we will represent conjunctive combination functions via separate arrows. Generally, the number of incoming arrows to the team outcome E shows the minimal number of agents who need to succeed in their individual tasks in order for the team to win.



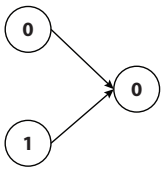
(a) Conjunctive Structure



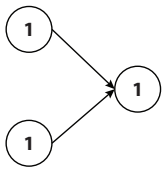
(b) Both fail, $Resp(A = 0, E = 0) = \frac{1}{2}$



(c) A succeeds, $Resp(A = 1, E = 0) = 0$



(d) B succeeds, $Resp(A = 0, E = 0) = 1$



(e) Both succeed, $Resp(A = 1, E = 1) = 1$

Figure 3.7: Conjunctive causal structure (top) with the four possible situations that can occur (bottom).

We will again focus on how much responsibility A receives for the group outcome in the four possible situations according to the structural model. In the situation in which both players failed in their individual tasks (Figure 3.7b), A has a responsibility of $\frac{1}{2}$. A is not pivotal in the actual situation. One change (changing B from *fail* to *succeed*, $N = 1$) is necessary to generate a situation in which A would have been pivotal. In the situation in which A succeeds (Figure 3.7c), A has 0 responsibility for the loss according to the same reasoning as outlined above – there is no possible contingency that would make $A =$ *succeeds* pivotal for the loss. In the situation in which B succeeds, A's responsibility for the loss is 1 (Figure 3.7d). Finally, in the situation in which both A and B succeed, A is pivotal and hence A's responsibility for the win is 1 (Figure 3.7e).

an *XOR* situation, A (and B) are fully responsible for the outcome in each of the possible situations.

To sum up, in a disjunctive structure, we see that according to the structural model, responsibility is shared between the players for positive outcomes (i.e. $A = B = 0.5$). For negative outcomes, in contrast, each player who failed receives full responsibility (i.e. $A = B = 1$). The opposite is true for conjunctive structures. Here the structural model assigns a responsibility of 0.5 to A when both failed and a responsibility of 1 when both succeeded. Table 3.1 summarises the predictions of how much responsibility A has for the different possible outcomes in the disjunctive and conjunctive setting.

Table 3.1: Responsibility of A for the disjunctive and conjunctive structures.

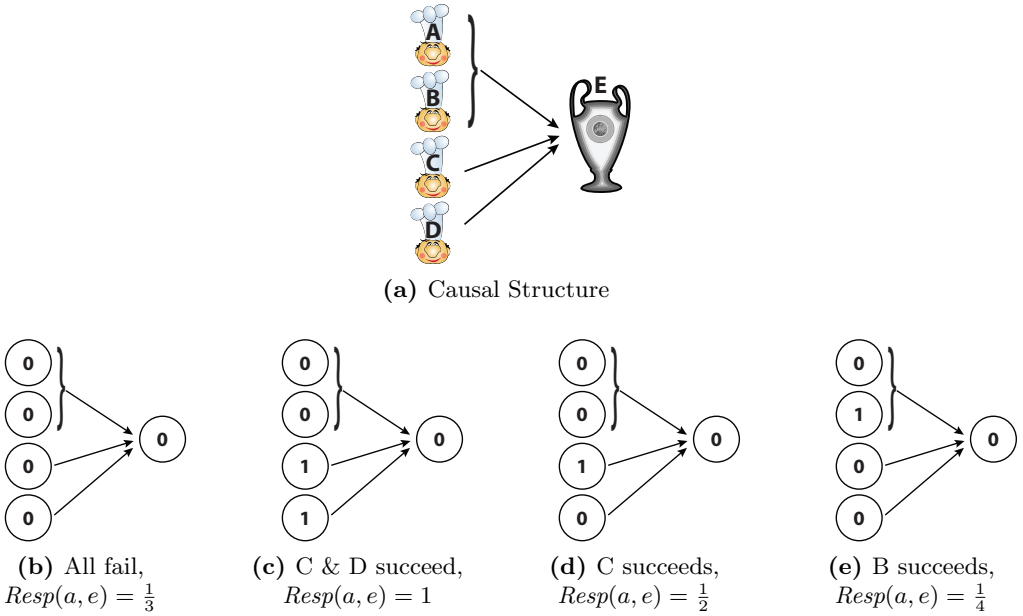
	Both fail	A succeeds	B succeeds	Both succeed
disjunctive	1	1	0	0.5
conjunctive	0.5	0	1	1

Note: Grey cells show A's responsibility for losses and white cells show A's responsibility for wins.

A slightly more complex example of how the model works As we have seen, the structural model makes precise quantitative predictions about how much responsibility agents carry for positive or negative outcomes in conjunctive and disjunctive causal structures. The structural model not only makes predictions for these simple causal structures, but also for more complex structures. Indeed, the structural model makes clear quantitative predictions for all situations that can be modelled in terms of binary variables that are related via deterministic causal links.

Consider the following scenario: as part of the infamous UCL Psychology Christmas Show, the organisers have put together a little cooking challenge. Professor A, B, C and D have been chosen to compete and are randomly assigned to different tasks. Professor A and B have to prepare starters independently from each other. Professor C prepares the main and Professor D the dessert. In order for the group to win the challenge (and secure everlasting fame) they have to cook a successful meal which includes a successful starter, main and dessert. Hence, both C (main) and D (dessert) and at least one out of A or B (starters) have to convince the tough judges with their cooking creations (thus, $E = min(max(A, B), C, D)$). Figure 3.8 shows the causal structure together with four of the possible situations that could happen in the challenge.

Let us focus on A again. In the situation in which all fail (Figure 3.8b), two changes are required to make A pivotal (i.e. changing C and D from having failed to having succeeded). Hence, A's responsibility for the group's loss is $\frac{1}{3}$. In the situation in which both C and D succeeded (Figure 3.8c), A is pivotal and hence has a responsibility of 1. In the situation in which C succeeded (Figure 3.8d), only one change is needed to make A pivotal and hence A's responsibility is $\frac{1}{2}$. Finally, in the situation in which B succeeded (Figure 3.8e), three changes would be necessary in order to make A pivotal. B needs to be changed from having succeeded to having failed and C and D need to



(a) Causal Structure

(b) All fail, $Resp(a, e) = \frac{1}{3}$ **(c)** C & D succeed, $Resp(a, e) = 1$ **(d)** C succeeds, $Resp(a, e) = \frac{1}{2}$ **(e)** B succeeds, $Resp(a, e) = \frac{1}{4}$

Figure 3.8: Causal structure of the cooking scenario (top). Some of the possible situations that can occur (bottom). *Note*: 0 = failed in their task, 1 = succeeded in their task.

be changed from having failed to having succeeded. Thus, A's responsibility in this situation is $\frac{1}{4}$.

We see that A's responsibility changes as a function of how many of the other players succeeded in their task. Furthermore, the relationship between A and the other players in the causal structure is crucial. Hence, A's responsibility for a negative outcome does not necessarily increase when fewer people failed in their task as a simple diffusion of responsibility model would suggest (Darley & Latané, 1968). If B succeeded, A is *less* responsible for the negative outcome than when noone succeeded. In contrast, if C (or D) succeeded, A's responsibility increases. Given the causal structure of the task, B's success moves A further away from pivotality whereas C's success brings A closer to pivotality. The importance of the underlying causal structure for how responsibility is attributed will be discussed in greater detail in Chapter 4.

3.2.1.2 Ruling out *non-blameworthy* causes

In the previous section, we have seen how the structural model of responsibility attribution allows for individual causes to have partial responsibility for an overdetermined outcome. It thus rules in causes as responsible that are ruled out by a simple counterfactual model. However, as outlined above, we saw that the simple *but for* test is both too exclusive and too inclusive. I will now briefly discuss in what way the structural model is also more exclusive.

Chockler and Halpern (2004) distinguish between the notions of *responsibility* and *blame*. *Responsibility* is an objective notion in that it merely depends on the values

of the variables as realised in the actual world. *Blame*, in contrast, is subjective and relative to the epistemic states of the agents in the situation. Consider the firing squad example once more in which two agents A and B are about to bring about the death of E in a disjunctive fashion. However, this time the situation is different in that A and B know that only one of their guns (selected at random) carries a live bullet. That is, for one of them, pulling the trigger does not affect whether E will die or not, whereas the other one will cause E's death with certainty. Imagine that both A and B shoot and that E dies. As it turned out, A's gun carried the live bullet while B's gun was empty.

According to Chockler and Halpern's (2004), A is the sole actual cause of E's death. He is pivotal and thus carries a responsibility of 1. However, we might say that it seems unfair that A should carry all the responsibility and B is let off the hook, given that both had the same *a priori* chance of bringing about E's death by their actions. Chockler and Halpern (2004) share this intuition and develop a different concept (which they call *blame*) that is relative to the epistemic states of the agents. In short, blame is defined as an agent's expected degree of responsibility. We need to consider all possible situations that can result from an agent's action, determine how much responsibility the agent would have for the outcome in each of these situations, and then average over these responsibility values to arrive at the person's blame. Hence, an agent's blame for a particular outcome is given by

$$\mathit{blame}(A, E = e) = \sum_{i=1}^{k} \mathit{responsibility}(a, e)_i \times \mathit{probability}(a, e)_i, \tag{3.5}$$

where i denotes each situation that the person considers and k the number of situations.

Let us consider how much blame A is predicted to receive in the Russian-Roulette-style situation just described. There are two possible situations (given that both of them decide to shoot): either A carries the live bullet or B does. In one of these situations, A's responsibility is 1 (if he carries the bullet) while in the other situation, his responsibility is 0 (if he doesn't carry the bullet). Averaging over these two possible situations, we get that A's blame for E's death is $\frac{1}{2}$. Although B's action did not cause E's death and B has no responsibility for the outcome, B is predicted to receive a blame of $\frac{1}{2}$ as well. We see that, according to Chockler and Halpern's (2004) definition of responsibility and blame, the two concepts can be strongly dissociated. If a person considers a situation extremely unlikely in which she will be responsible for the outcome, but this situation happens to come about, she will receive little blame but high responsibility. In contrast, if a person thinks that it is very likely that she will be responsible for an anticipated negative outcome but in fact it turns out that she isn't, her blame will be high without her being responsible for the outcome.

Given that the concept of blame is dependent on the epistemic beliefs of an agent, two agents can receive unequal blame for the same outcome in the same situation.

Consider that A and B were tricked: although they were told that only one of their guns would carry a live bullet, in fact, both of their guns did. While A considered this possibility and assigned some probability to it, B did not consider this situation at all. Furthermore, B had a hunch that A's gun was more likely to carry the bullet than his. Hence, A's and B's blame for the negative outcome would be quite different (see Table 3.2).

Table 3.2: Blame of agents A and B in the execution scenario. *Note*: Probability() = probability that each agent assigns to the different situations.

Bullet?	Probability(A)	Responsibility(A)	Probability(B)	Responsibility(B)
A	0.4	1	0.8	0
B	0.4	0	0.2	1
both	0.2	0.50	0	0.5
	blame(A) = 0.5		blame(B) = 0.2	

With this concept at hand, we can now see how the structural model rules out events (at least from blame) which would be ruled in by a simple counterfactual model. Consider, for example, the long chain of events in which A's selling a weapon led to E being shot described above (see Figure 3.3a). While A turns out to be *responsible* for the outcome, he would be predicted to receive little *blame*. The probability that this particular outcome would have resulted from A's action is very low – the negative outcome was not foreseeable. For example, we can imagine that A assigned most of the probability that the weapon would be used for hunting deer in the woods and not for killing people. Furthermore, there was the potential for other events to intervene and break the chain of causation.[11]

Consider the situation in which a person A throws a lighted cigarette into a bush in a forest (cf. Hart & Honoré, 1959/1985). Just as the small fire in the bush is about to go out, B comes along and pours oil over it. As a result, there is a massive forest fire. As Chockler and Halpern (2004) note, whether or not A is a cause of the fire depends on the causal model of the situation. If B had started the fire independently of what A did, then A is not the cause and hence has no responsibility. In legal terms, B's action would have broken the causal chain from A's action to the negative outcome.

However, if B only succeeded in causing the forest fire *because* A's cigarette had set the bush on fire (say, for example, that B did not have any tools on him to start a fire), then the responsibility of both A and B would be 1 (assuming A's action in itself would not have been sufficient to start the forest fire). Let us now consider different possible epistemic states that A might have had. If, for example, A happened to know that B

[11] Woodward (2006) makes a similar point regarding people's judgments of actual causation. Accordingly, people are reluctant to call an event C the cause of another event E when the relationship between C and E is *sensitive*. The causal relationship is *sensitive*, when the probability of the two relevant counterfactuals (i) "If C were to happen, E would happen" and (ii) "If C had not happened, E would not have happened" is low.

is a notorious firebug, saw that he followed him and knew that he would not be able to resist setting fire, then A is highly to blame. Here, the negative consequences of his action were foreseeable. Yet, if A had no idea that throwing the cigarette in the bush would have had such negative consequences, A's blame would be low.

3.2.1.3 Some shortcomings of the structural model of responsibility attribution

In this section, I will outline some of the shortcomings of the structural model of responsibility attribution. The empirical tests of the model described in Chapter 4 will address some of these concerns and motivate important theoretical extensions to the structural model.

What is a *minimal change*? We have seen above that Chockler and Halpern's (2004) model defines responsibility in terms of the *minimal number of changes* that have to be made to the actual situation to generate a possible situation in which the variable under interest is pivotal for the outcome. The minimal number of changes is defined in terms of the *number of variables* in the network whose value needs to be changed. However, this notion is problematic.

Consider a situation in which different variables have different prior probabilities of occurring. For example, imagine that a medium-skilled pool player A plays in a team with either a highly-skilled player B_{good} or a player with a low skill level B_{bad}. Let's say that in both situations, B misses a very easy shot before the other team wins the game. Now in order to consider how responsible A was for the outcome, we need to consider the situation in which B hadn't missed the easy shot. In terms of the structural model, we need to change the value of one variable, B, from missing to hitting. However, it seems that there is a difference in how difficult this change is to make between the situation in which A plays with B_{good} versus B_{bad}. It seems less of a stretch to change B_{good} from *missed* to *hit* compared to doing the same change for B_{bad}. That is, the partner's prior performance affects how easily a possible change of a variable's value can be imagined (see Kahneman & Miller, 1986). However, the notion of a change as used by Chockler and Halpern (2004) is insensitive to differences in the prior probabilities of variables.

This example illustrates that changing a minimal number of variables might not necessarily be the most 'minimal' way in order to generate a close possible world. A possible world in which several variables were changed whose values in the actual world were unexpected might be closer to the actual world than a possible world in which fewer variables were changed who realised their expected values.

Chockler and Halpern (2004) discuss another potential problem of their account. Compare the situation in which a vote between two candidates A and B is 2 – 0 versus 100 – 98. According to the structural model, a person who voted for A is equally

responsible for the outcome in both situations. However, it feels like each person who voted for A is less responsible for the outcome in the situation in which there were 99 others who also voted for the same candidate. This again shows that minimal changes are not always the same: making one change in the situation with more voters seems more minimal than making a change in the situation with fewer voters.[12]

What happens when variables are not binary? Thus far, we have constrained our discussion to situations in which we can model the relevant events in terms of binary variables. However, there are many situations in the world which cannot readily be modelled in black and white. Consider the situation in which a local library can only be saved if more than £1,000 is donated. As it turns out, A donates £1 and B donates £1,000. Ignoring possible differences between the donors's wealth or motives, there should be a close correspondence between how much each person donated and how responsible they are for saving the library.[13] However, in terms of the structural model both A and B are pivotal and thus fully responsible for the outcome. If either of the two had not donated, the library would not have been saved.

Chockler and Halpern (2004) discuss that this problem can be resolved by assigning weights to variables. Thus, A's responsibility in the situation in which the outcome was overdetermined would be $\mathit{Resp}(A = £1, E > £1,000) = \frac{£1}{£1+£1,000} \approx 0.001$, whereas B's responsibility would be $\mathit{Resp}(B = £1,000, E > £1,000) = \frac{£1,000}{£1,000+£1,000} \approx 0.999$. However, this simple weighting strategy will not always work. For example, imagine that a clever son proposes that each member of the family should vote for where to go in their next holidays. The son happens to know that his father would like to go on a fishing holiday just like he does. However, his mother would rather go on a beach holiday. In order to make things fair, the son proposes that the mother should have five votes, the father four and the son only two votes. Each person has to vote for either of the two options and whichever option gets the majority of votes is chosen. As it turns out, both the father and the son vote for the fishing trip whereas the mother voted for the beach trip. Should the father be held more responsible for this result than the son? It seems like weighting responsibility is wrong here: in fact, the votes were chosen in such a way that each person had an equal influence over the outcome of the election. Any combination of two voters always yields a majority (independent of the assigned weigths).[14]

[12]Note that Chockler and Halpern's (2004) concept of *blame* captures the difference between a voter for A in the two situations. Assuming that each voter has an equal probability of voting for each candidate, a voter for A is more likely to carry more responsibility for the outcome in the situation which involves less voters (see also Kerr, 1989; Rapoport, 1985, 1987).

[13]As he looked up, Jesus saw the rich putting their gifts into the temple treasury. He also saw a poor widow put in two very small copper coins. "I tell you the truth," he said, "this poor widow has put in more than all the others. All these people gave their gifts out of their wealth; but she out of her poverty put in all she had to live on." (Luke, 21: 1 – 4).

[14]That the number of votes does not directly translate into how much influence a person (or state) has over the outcome of an election is a well-known fact in political science (see, e.g. Banzhaf, 1964;

An even deeper problem than the problem about whether individual contributions to a joint outcome should be weighted, concerns the *evaluation of counterfactuals* in situations in which variables leave the safe haven of a binary world. For binary variables, a counterfactual is clearly defined: if a person succeeded, we consider what would have happened if the person had failed. However, as soon as variables are continuous, there are an infinite number of counterfactuals that could potentially be considered. For example, consider the counterfactual "If I had spent more time on my thesis it would have turned out better". Well, how much more time would I have needed to spent? How much better would it have been as a result? In order to answer these questions, we need to know about the exact functional dependence between the different variables involved. This severely aggravates the problem of what counts as a minimal change described above.

How is the model constructed? Responsibility as defined by Chockler and Halpern (2004) is model-dependent. That is, if two people have constructed a different causal model of a particular situation, then their responsibility attributions are predicted to be different. Thus, the structural approach suggests that disagreements about how much responsibility a person should carry for the outcome should be resolved once people have agreed upon the underlying causal model that generated the outcome. However, sometimes the model-dependence appears to go too far.

Chockler and Halpern (2004) give one example for how model construction influences attributions of responsibility. Again, consider a voting scenario in which candidate A wins against candidate B 11 – 0. Now, in one model we represent each voter X_i in terms of an individual variable. In this model, the responsibility of X_1 for the outcome is $\frac{1}{6}$. However, let's consider another model in which we only use two variables: one for X_1 and another one Y for the rest of voters, now with a range from 0 to 10. In this model, X_1's responsibility reduces to $\frac{1}{2}$ since only one variable needs to be changed to make X_1 pivotal. Of course, this problem is related to both the problem of minimal changes as well as the problem of having non-binary variables.

Recently, Livengood (2011) has shown that current models of actual causation such as the model developed by Halpern and Pearl (2005) on which Chockler and Halpern's (2004) structural model is based struggle with quite simple voting scenarios. Livengood (2011) shows that for elections in which abstentions are allowed, current theories of actual causation (Hall, 2007; Halpern & Pearl, 2005; Hitchcock, 2001b; Woodward, 2003) rule in *every* abstention as a cause of the election outcome irrespective of how close it was. Even more worrying: for elections which involve more than two candidates, all of the considered theories count every vote (no matter for which candidate) as a cause of the winning candidate's victory. These observations suggest, that current models of actual causation appear to be overly sensitive to structural changes that should not

Felsenthal & Machover, 2004; Shapley & Shubik, 1954).

really make a difference – model construction becomes an art (Halpern & Hitchcock, 2011).

What about intentions? In the previous chapter, we have seen the crucial importance of intentions for attributions of responsibility and blame. However, whether or not an action was performed intentionally does not affect how responsible it was for an outcome in Chockler and Halpern's (2004) account. It does potentially affect attributions of blame but only to the extent that information about intentions change one's expectations about what situations are likely to result from a person's action (see Brewer, 1977; Spellman, 1997). Providing a formal account of how intentions can be modelled in causal Bayes nets remains a challenge for future research (Baker et al., 2009; Goodman et al., 2009; Ullman et al., 2009).

3.3 Conclusion

In this chapter, I have provided an overview of recent work that has emphasised the close relationship between causality, counterfactuals, responsibility and blame. I have shown how these concepts can be given a firm formal interpretation in terms of the causal Bayes net framework. Furthermore, we have seen that causal Bayes nets can be used to determine which events count as actual causes of an outcome of interest. While the *but for* test works as a test for actual causation in many situations, the simple counterfactual criterion is both too exclusive and too inclusive. A relaxed notion of counterfactual dependence is required to allow for individual events to be causes in situations of overdetermination. An event can still count as an actual cause when it made no difference to the outcome in the actual situation but when it would have made a difference in another possible situation. I have introduced the structural model of responsibility attribution (Chockler & Halpern, 2004) which takes this relaxed notion of counterfactual dependence as a starting point. A person's responsibility is a function of the distance to a world in which the person's contribution would have made a difference to the outcome. I have provided a few examples of how the model works and have discussed some of its shortcomings. In the next chapter, we will see how well the structural model accounts for people's attributions of responsibility in different experimental setups.

Chapter 4

Causal Structure and Responsibility

> With great power comes great responsibility.
>
> – **Spiderman**

The previous chapters have equipped us with the theoretical background of the responsibility attribution literature (Chapter 2) and the formal tools (Chapter 3) we need for modelling attributions of responsibility. Now it's time to dig in and get our hands dirty. In this chapter, I will discuss several experiments that were conducted in order to investigate how people attribute responsibility to individuals within a group.

One of the central features of the structural model (Chockler & Halpern, 2004) is that it can deal with situations of overdetermination. We have seen that previous formal models of responsibility attribution are based on the simple notion of counterfactual dependence and that they are hence forced to conclude that each individual event bears *no responsibility* as soon as the outcome is overdetermined (see Brewer, 1977; Fincham & Jaspars, 1983; Petrocelli et al., 2011; Spellman, 1997). Just to reiterate, this is because in such situations, the outcome would have occurred *even if* the individual causal event had been different. Thus, in a simple sense, what one person did made no difference to the outcome.

Due to the examples I have used in the previous chapter, such as firing squads shooting a prisoner, you might have the impression that situations of overdetermination are like an endangered species: rare and difficult to spot in the wild. Why, you might wonder, devote so much effort to a problem that hardly ever occurs? However, I would argue that quite the contrary is the case: situations of overdetermination are ubiquitous! In fact, no day passes in which we do not encounter a host of situations of overdetermination – however, we might not always spot them (see Schaffer, 2003). Let me mention

just a few examples.

We often do things for a reason. For example, I am writing this thesis because I would like to get a PhD. But – hold on – I am also writing this thesis because that's what I've set myself as a task for today, because I would like to get a PostDoc position afterwards at a good institution, because I want my family and my supervisors to be proud of me, because ... It seems like we often do things for *many* reasons! However, which out of the many reasons is *the one* that causes me to sit here and wonder about overdetermined effects? Would I not sit here if I lacked the reason of wanting to make my supervisors proud? If this were so, would it mean that this particular reason is not responsible at all for my sitting here?

As this example illustrates, our actions are often overdetermined through many reasons (cf. Mele, 1997). Indeed, this is part of the problem of what makes inferring another person's reasons for why he or she behaved in a certain way a difficult task: there is a *many-to-many mapping* from reasons to actions. There are many reasons for any particular action and many actions that can result from one particular reason. Attribution theory, as discussed in Chapter 2, has detailed some of the strategies on which people rely to find their way through the maze that connects actions and reasons.

Let's move on to another example: Mackie (1974) has argued that many effects in the world are *quantitatively overdetermined*. If he is right, then the few paragraphs you've read in this chapter thus far are due to more than 3,000 instances of overdetermination. Mackie (1974) illustrates his point via the proverb "To use a sledge-hammer to crack a nut".[1] Of course, the point is that the full blow of a sledge-hammer is not necessary to flatten a nut. A smaller hammer would have been sufficient to do the job. In other words, we could have removed parts of the sledge-hammer, and the nut would still have been flattened. Analogoulsy, each time I am pressing a key on my keyboard, I overdetermine the effect of the key being pressed – maybe, a slightly softer press (or a slightly smaller finger) would have done the job.

These two examples illustrate that one doesn't need many people to create situations of overdetermination. However, the focus of the experiments in this chapter will be on cases in which outcomes were overdetermined through the actions of several people. Because funding for my PhD was generously provided by the AXA Research Fund, it might be appropriate to use an insurance-related example as an illustration. Imagine that you are in the process of building a small family house in the outskirts of London. In order to fix the roof, you need wood for the framing and tiles to put on top. You've ordered the tiles and wood from different contractors. You can only start building the roof if you have both the wood and the tiles. However, as it turns out, the tiles are delivered three weeks late. Even worse, the wood is delivered six weeks late. Thus, the building process is considerably delayed which results in significant additional costs. Of

[1] The German equivalent of this proverb is: "Shooting at sparrows with canons."

course, you demand financial compensation from the two contractors. However, here is their defence: *Mr Tile* says "I am not going to pay any compensation. Even if we had delivered the tiles on time, this would have made no difference to the delay whatsoever because the wood was six weeks late." to which *Mr Wood* replies: "Fair enough, I agree that we have to pay compensation for a three-week delay. However, we won't pay for six weeks! Even if we had delivered on time, there would still have been a three-week delay because of the late delivery of the tiles."[2] The first three weeks of the delay are overdetermined and each party refuses to pay compensation for this period. Imagine you are the judge: How much compensation should each party pay?

There are many other situations in which the actions of individual people combine to bring about an outcome in an overdetermined fashion. Some of the examples on the larger scale include the recent financial crisis, the looting during the riots in London, 2011 or the effects of each person's actions on global warming. In each of these situations, part of the reason for why individuals are reluctant to take responsibility for the outcomes of their actions is because they feel that their individual contributions didn't make a difference (see Kerr, 1996; Kerr & Kaufman-Gilliland, 1997). The bank will go bust even if I don't try to squeeze the most money out of it. The plane would fly even if I were to take a train instead. Someone else would go and take this TV for free if I don't. These situations exhibit the classic *tragedy of the commons* structure: what is best individually creates a terrible outcome collectively (Hardin, 1968). There are many more smaller-scale examples in which people work together in groups, such as team sports or task forces, that bear the potential of creating situations of overdetermination. For example, I just read in our local newspaper about the result of a highly overdetermined outcome in a football game (11 – 0).

These examples have shown that situations of overdetermination are not only abstract problem cases of particular interest to philosophers who enjoy arguing against counterfactual theories of causation but that they are an inherent part of our everyday life. Understanding how people make attributions in such cases is both of theoretical and practical interest. The central question of this chapter will be the following: How much is each individual group member responsible for the outcome that their group has brought about as a collective? As the title of this chapter suggests, I will argue that any adequate answer to this questions has to start from a careful analysis of the causal structure of the group context.

The remainder of this chapter is organised in the following way: I will first review some of the previous philosophical and psychological work that has investigated responsibility attributions in groups. I will then report an experiment which establishes the

[2]Of course, in reality, one would have an individual insurance with each contractor which would resolve this problem. However, there are indeed some legal cases reported in Hart and Honoré (1959/1985) which are of the same nature and in which people have avoided paying via the counterfactual reasoning illustrated in the example.

importance of causal structure for attributions of responsibility. The causal structure determines, for example, how the effects of each person's actions combine to determine the outcome of the group. We will see that the underlying causal structure affects the extent to which individuals in a group are held responsible (Gerstenberg & Lagnado, 2010). The results of this experiment will motivate a further set of studies in which we look at attributions of responsibility in asymmetric group structures (Zultan et al., 2012). In each of the reported experiments, we find support for some of the core predictions by the structural model of responsibility attribution (Chockler & Halpern, 2004): people tend to receive less responsibility the further they are away from being pivotal. However, the empirical results will also motivate extensions to the model. Most importantly, we will see that attributions of responsibility are not only influenced by whether a person's contribution made a difference to the outcome (or by how close the person was to making a difference) but also by how critical an individual's contribution was perceived for the group outcome in the first place (Lagnado et al., accepted).

4.0.1 General features of the experiments

Before diving straight into the experiments, I will make some remarks about the general features that all the experiments in this chapter have in common. As discussed in Chapter 2 there are several factors that influence responsibility attributions such as an agent's causal contribution to the outcome as well as their mental states. In this chapter, we will be interested in the influence that an agent's causal status has on their responsibility for an outcome. The experiments in this chapter are all set in an achievement context (cf. Weiner, 1995). Several individual players form teams and the question is how much each player in a team is to blame for the team's loss or to credit for the team's win. The mental states of the individuals in the group, such as their intentions or beliefs, are not varied.[3] All features of the setup are common knowledge and it can be assumed that each person will try their best to be successful. In contrast to the traditional social psychology approach, which relies almost exclusively on scenario-based research, the experiments in this chapter will be based on simple group games. This research strategy enables us to generate patterns of observations for which we can derive precise quantitative predictions from our responsibility attribution models. We can then test whether the models predict people's attributions accurately and refine the models if they don't.

Except for the first experiment, our participants will make attributions of responsibility from the perspective of an external observer or judge. We have seen in Chapter 2 that actors and observers sometimes reach quite different attributional verdicts (Jones & Nisbett, 1971; Malle, 2006). The reason why we mostly focus on the attributions of

[3]In Chapter 5, we will see how attributions of responsibility are influenced by manipulations of the agents' mental states.

observers is a methodological one: when people form part of a group they influence the generation of the data. This renders the comparison of different participants' behaviour problematic. A person who performed well in a task will arguably differ in their attributions from a person whose performance was poor. While many social psychology studies have relied on deceptive methods, such as false feedback about performance, in order to facilitate comparison between participants we did not want to follow suit. In fact, none of the experiments reported in this thesis involve any form of deception. Looking at the attributions of an uninvolved observer serves as a good starting point. Future research will need to investigate how actors differ from observers in the paradigms that we have developed.

The problem of attributing responsibility is aggravated by the fact that causes are often difficult to identify and, if identified, hardly ever bring about effects in complete isolation. In all experiments in this chapter, we have made the task for participants somewhat easier by having selected the relevant causes for them. Participants were provided with complete information about the causal structure and the functional relationships between the different variables. You might think that this renders the task of assigning responsibility trivial. However, the following thought experiment might help to illustrate how difficult the problem of responsibility attribution actually is.

Assume you are given a full-length video tape of a Premier League football match between Arsenal and Chelsea. You are given access to the most powerful computers equipped with state-of-the-art machine learning software, motion capture, face detection, etc. Furthermore, you have infinite time for completing your task. The task, as you can already imagine, is to assign a value of responsibility to each player for their team's result. Despite the fact that you know everything about the rules of the game, the players in each team (in fact, you have access to the complete history of each player from birth onwards including a fine-grained psychoanalytic analysis) and the events that took place in the stadium (there were cameras everywhere and you can slow the clips down as much as you like), I will argue that you will still find the task extremely difficult.

Interestingly, there is now a commercially available solution to the problem that I have just described – the Capello Index (CapelloIndex.com, 2012). Here is a direct quote from their homepage explaining how the index works:

> "Using a scoring system which takes account of every key event that occurs during the course of a match, the Capello Index has a unique formula that measures a player's contribution from both a quantitative and qualitative perspective. ... Football is, by definition, an unpredictable sport, an *inexact science*, which therefore explains why no one has yet managed to create a technology capable of codifying this game. Having been able to have a look at the sport through the eyes of a manager like Fabio, for me everything now seems *different and understandable* and I hope *fans around the*

> *world* will feel the same when they learn more about the Capello Index." (CapelloIndex.com, 2012, emphasis added)

The fact that the Capello Index has been heavily criticised (not only because many players of the English football squad received quite low ratings) can be taken as evidence that the problem is far from being solved – there is still considerable disagreement. Nevertheless, I would like to follow up on some of the points that have been raised in the quote above: I hope that *readers around the world* will see this chapter as a first step to understanding the practice of responsibility attribution from an *exact* scientific perspective and that things might look *different and understandable* afterwards.

4.1 Responsibility in Groups (Gerstenberg & Lagnado, 2010)

The football scenario described above served as an example to illustrate the problem of attributing responsibility between multiple people in a group. I will first summarise some of the philosophical and psychological research that has investigated attributions of responsibility within groups before discussing our study.

4.1.1 Can a group be *collectively* responsible?

For philosophers, a fundamental question with respect to collective responsibility is whether it is reducible to individual responsibility or better understood as an emergent property that cannot be explained in terms of its constituting components. *Individualists* with respect to collective responsibility believe that it can be analysed in individual terms (e.g. Lewis, 1948; Mäkelä, 2007; Narveson, 2002; Sverdlik, 1987; Weber, 1914/1978) whereas *collectivists* believe that the synthesis of several individuals into a collective creates new phenomena that are irreducible (e.g. Cooper, 1968; Feinberg, 1968; French, 1984; Gilbert, 2006; Mathiesen, 2006; May, 1993; Pettit, 2007; Sheehy, 2003). In short, the question is whether the whole is more (collectivists) or equal to (individualists) the sum of its parts (see Smiley, 2011, for a review of the philosophical literature). The following two quotes nicely illustrate the two positions: In the right corner, representing the collectivists, we have Thomas Hobbes (1982/1651) who has claimed that the people who form a plural subject are "reducing all their Wills ... unto one Will" and thus create "a real Unitie of them all" (p. 227). And in the left corner, representing the individualists, we have Edward, First Baron Thurlow who countered: "Did you ever expect a corporation to have conscience, when it has no soul to be damned, and no body to be kicked?" (quoted in Coffee, 1981, p. 386)

With respect to the question of collective responsibility, the philosophical debate concentrates on the concepts of *action* and *intention*. Can a group have a collective

intention and act collectively? Individualists maintain that actions and intentions are properly understood to arise on the level of individuals. For example, can we hold a firing squad as a whole morally responsible for a prisoner's death? Individualists argue that the firing squad (as a whole) neither acts nor intends and thus the answer should be negative. Groups do not meet the criteria that are required for an agent to be held responsible (e.g. they do not make decisions). Furthermore, they argue that attributing collective responsibility would often be unfair (or even, 'barbarous' as Lewis, 1948, claims): imagine that one of the marksman had actually refrained from shooting but we still continued to hold the group responsible for the outcome. There is a strong intuition that such an attribution would be unfair towards the marksman who had the courage not to shoot. Similar arguments have been raised with respect to whether a country can be collectively held responsible. For example, should the Germans (as a nation) be held collectively responsible for the terrors of the Nazi regime (Thompson, 2006). Or should the Americans (as a nation) be held responsible for the Vietnam war?

Some collectivists have taken people's linguistic practice of blaming collectives as evidence in favour of their claims. Recently, psychological studies have supported what had mostly been anecdotal evidence thus far: people are often willing to attribute mental states to groups. Indeed, Waytz and Young (2012) have argued that there is a trade-off in that the more people attribute a 'group mind' to a collective, the less they perceive each person in the group to have an 'individual mind'. Waytz and Young argue further that this has important implications about how people attribute responsibility to the group and group members. In a similar vein, people can also experience guilt for injustices brought about by a collective they belong to despite the fact that they have not causally contributed in any way to that outcome. While many philosophers argue that such feelings of guilt are inappropriate in so far as guilt should only be felt with in relation to one's own actions, Gilbert (1997) argues that these feelings of guilt can be justified within a collectivist conception of responsibility. Thus, in analogy to how Strawson (2008) has argued that our our emotional responses towards individuals justify that we hold them responsible despite the possibility that there is no freedom of will, some collectivists (e.g Gilbert, 1997; Tollefsen, 2006) have claimed that responsibility attributions to collectives are justified by the emotional sentiments we have towards them.

Assuming that it is possible to hold groups (morally) responsible in a principled way, a natural follow-up question concerns *what* groups are appropriate targets of responsibility. Can any odd group be a suitable bearer of collective responsibility or do certain conditions have to be met in order for a group to be responsible? Collectivists have proposed different criteria for distinguishing groups that could potentially be held responsible from the rest. French (1984), for example, has focused on organisations as bearers of collective responsibility and has argued that the fact that these groups have organised decision-making procedures renders them potential candidates for responsibility. Others have argued that not the organised decision-making procedures are crucial

(which would rule out many groups) but that the group is comprised of individuals who share attitudes, have a common interest and are jointly committed to each other in their actions (Feinberg, 1968; Gilbert, 2006; May, 1993).

4.1.2 How is (collective) responsibility distributed between group members?

The second central question in the context of collective responsibility concerns the question of how the responsibility of the group is distributed between the individuals within the group. While it might be relatively straightforward to claim that the firing squad is (at least causally) responsible for the prisoner's death, the question of how much each of the individual marksmen should be held responsible is far from trivial. The problem of attributing responsibility to individuals is often aggravated by the fact that the individual contributions to the group outcome might be difficult to discern (Feinberg, 1968). While the individual contributions are relatively straightforward in the firing squad example, attributing responsibility to individual football players for their team's loss or win is more difficult because of complex interactions between the players, differences in what role each player occupies, the influence of chance factors, etc.

There has been considerable disagreement about whether individuals can be held partially responsible for an outcome that is indivisible such as a person's death. Zimmerman (1985), for example, has argued that each individual in a group is fully responsible for the outcome that was brought about by the group. Others, in contrast, have argued that responsibility between individuals can be shared even in situations in which the outcome cannot be divided (Cohen, 1981; Narveson, 2002). Thus, each of the individual marksmen could be attributed a fraction of the total responsibility. This tension between whether an individual who acts in a group should be held fully responsible or only partially responsible for the outcome is also reflected in different standards adopted by criminal and tort law (see Hart & Honoré, 1959/1985). In criminal law, an individual's responsibility is not reduced for a crime that is jointly committed. In tort law, in contrast, degrees of responsibility are distributed among the individuals involved in proportion to each individual's fault.

In social psychology, *diffusion of responsibility* is a well-known phenomenon (Darley & Latané, 1968; Latané, 1981; Latané & Nida, 1981). It refers to the fact that individuals are less likely to take responsibility for an action (or an omission to act) when others are present as well. Research in this area has been sparked by the murder of Kitty Genovese in 1964. At the time, an article was published in the New York Times with the headline "37 Who Saw Murder Didn't Call the Police" which resulted in much public debate.[4] In Darley and Latané's (1968) investigation of the *bystander effect*, the

[4]Later investigations established that many of the reported facts in the original article were wrong (e.g. nobody witnessed the attacks in their entirety, most only heard faint noises and did not think that something was seriously wrong; see Manning, Levine, & Collins, 2007).

probability that a participant intervened in order to help a confederate decreased from 85% when alone to 31% in the presence of four bystanders.

Darley and Latané's (1968)'s results support the *diffusion of responsibility* explanation only indirectly via the probability that an individual will help.[5] However, there has also been direct support that responsibility does in fact decrease with an increased group size both when judged from an insider's or an outsider's perspective (see Fischer et al., 2011, for a meta-review). Mynatt and Sherman (1975) found that individuals perceive themselves to be less responsible for negative outcomes when they acted as part of a group. Furthermore, Feldman and Rosen (1978) have shown that external observers hold individuals acting alone more responsible for a negative outcome than each individual member in a group. They found an inverse relationship between group size and attributed responsibility to each group member. Interestingly, they also examined the prison sentences of offenders in actual court proceedings. In contrast to the principle of joint enterprise in criminal law as discussed above, they found that offenders who acted alone were punished more severely than offenders who had acted in a group.

While experiments have shown that responsibility generally decreases with an increased group size (Feldman & Rosen, 1978; Mynatt & Sherman, 1975), researchers have also tried to investigate the exact functional form according to which responsibility decreases. For example, does responsibility decrease linearly, exponentially or in a power function as the group size increases? Do people have a tendency to normalise their responsibility allocations within a group such that an increased responsibility to one person implies a decreased responsibility to another person? Latané (1981) suggested a *psychosocial law* (following psychophysical laws; cf. Stevens, 1957) according to which the social impact I perceived by a person increases with some power t of the number of other people present N. Thus,

$$I = sN^t, \tag{4.1}$$

where s is a scaling constant and $t < 1$. Since t is assumed to be smaller than 1, the impact of each additional person is diminished. Latané (1981) has used this general law as a unifying explanation for diffusion of responsibility and many other social psychological effects. According to Equation 4.1, the largest drop in perceived responsibility is predicted to occur when a second person is present. With each additional person, the degree to which a person's perceived responsibility reduces is predicted to be smaller.

Teigen and Brun (2011) investigated systematically how individual responsibility changed as a function of the group size. Specifically, they were interested in whether people tend to adopt a *distributive view* according to which an overall amount of re-

[5] Alternative explanations for the bystander effect include that people in groups can sometimes be less likely to notice that something is wrong or that they are less likely to interpret the situation as one in which it is appropriate to help because the others are not helping either (Darley & Latané, 1968; Latané & Nida, 1981).

sponsibility is distributed between the individual group members or a *individualistic view* according to which individuals in the group are largely treated independently. Both views have some intuitive appeal: on the one hand, one could argue that there is indeed an overall responsibility for a particular outcome caused by the group that is to be divided between the individuals. If one group member acted very badly, this can exonerate another group member. On the other hand, it seems plausible that we would sometimes like to hold each person in a group very responsible (e.g. when something really bad happened) or not so responsible (when something of minor significance happened). A distributive view which enforces normalisation would prohibit a distinction between these two situations.

Previous research had found evidence for a *distributive view* of responsibility attribution in groups. However, some of these studies have enforced a distributive result methodologically by having asked participants to allocate a fixed amount of responsibility points between the group members (Feigenson, Park, & Salovey, 1997; Forsyth & Kelley, 1994; Forsyth, Zyzniewski, & Giammanco, 2002). Naturally, in such situations, individual responsibility decreases stronger with an increase of the 'group size' from one to two (e.g. from 100 to 50) compared to an increase from two to three (e.g. from 50 to 33). Other studies, in contrast, found support for the *individualistic view*. For example, Savitsky, Van Boven, Epley, and Wight (2005) found that individuals tended to take more than their 'fair share' of responsibility for positive outcomes when the other group members were conceptualised as 'the rest of the group'. In contrast, when the group was 'unpacked' (i.e. each individual group member listed), participants' tendency to indulge in self-serving attributions decreased. Again, however, participants were instructed to sum up the overall responsibility points to 100.

In order to have a fairer test of the *distributive* versus *individualistic* view, participants in Teigen and Brun's (2011) experiments, were asked to estimate responsibility in terms of percentage scores without being constrained to a fixed overall score. Participants read different vignettes in which the actions/decisions of groups with different sizes led to positive or negative outcomes. Teigen and Brun found that while participants tended to attribute an overall responsibility of 100% when groups consisted of two members, the overall responsibility exceeded 100% for larger groups with three or four members. More specifically, adding additional members to a group of two did not affect the responsibility attributed to these first two members.

Teigen and Brun propose the *singularity principle* as a potential explanation for this seemingly puzzling pattern of results. According to the *singularity principle* (Evans, 2006), people can only consider a single hypothesis (or mental model) at a time when evaluating evidence. Thus, people are assumed to evaluate the responsibility of each person separately without considering the actions of the others. In the two actor case this will look as if people distribute responsibility since "one actor's absence of responsibility will be sufficiently similar to another actor's presence of responsibility to make the

responsibility estimates appear additive and compensatory" (Teigen & Brun, 2011, p. 38). In contrast, if more than two actors are listed, several comparisons between actors would have to be made. The authors speculate that because of the potentially large number of pairwise comparisons, participants have to use a simpler strategy and not consider all members of the group. Instead, once the responsibility of a pair has been determined, people might consider the next person's responsibility without revising the responsibility that has already been assigned. Teigen and Brun (2011) admit that their explanation is rather speculative. Indeed, I would argue that Teigen and Brun as well as others who have looked at how responsibility diffuses in groups (e.g. Forsyth et al., 2002) have paid insufficient attention to the causal structure.

In Experiment 1, Teigen and Brun used three different scenarios. In the *job scenario*, Johanne is uncertain whether to take a job and asks her friend (group of friends) for advice. In the *article scenario*, Hans has written a paper with Grete (or with three other authors) whereby each co-author was in charge of a different section of the article. In the *relay scenario*, two teams consisting of two (or four athletes) compete against each other. While all three scenarios are identical in terms of the size of the group (which is manipulated to be either two or four), the different scenarios differ markedly in the ways in which individual contributions combine to yield the group outcome.

Note that in the *job scenario*, the other friends are not necessary for Johanne's decision which she could have made by herself. Maybe one positive answer by a friend would be sufficient to make her take the job. In the *article scenario*, in contrast, the contributions of each group member are necessary in order for the group of scientists to be successful. If any of them had failed to write their section, the article could not have been submitted. Finally, in the *relay scenario*, individual performances combine in an additive manner whereby faster athletes can compensate for slower athletes in the team. Interestingly, participants responsibility attributions were sensitive to these structural differences – a result which Teigen and Brun (2011) failed to report. In the *job scenario*, Johanne's friends received low individual responsibility ($M = 22.78\%$) for the outcome. In contrast, in the *article scenario*, the three co-authors each received high degrees of responsibility ($M = 52.7\%$) for the outcome. Finally, in the *relay scenario*, attributions were also high ($M = 52.58\%$).

These results show that participants were sensitive to the different ways in which individual contributions combined. Responsibility between the three friends in the *job scenario* diffused more than between the three authors or the three athletes. Thus, whether and how responsibility diffuses does not only depend on the number of people in the group but also on the causal relationships between the people in the group and how individual contributions combine to determine the group outcome. The dichotomy between the *individualistic view* and the *distributive view* does not exhaust the space of theoretical possibilities. In fact, I will show below that people's responsibility attributions are more sophisticated: whether attributions look 'distributive' or 'individualistic'

depends crucially on the way in which the contributions of the group members combine. According to such a view, it is possible that an individual will in fact be held more responsible when acting in a group compared to acting alone. If, for example, Tom's action strongly affects what other people in his group do, Tom could be blamed more for a negative outcome to which many others have contributed as well than a negative outcome that was brought about by Tom alone.

4.1.3 The influence of causal structure on attributions of responsibility

In his seminal book on group processes and productivity, Steiner (1972) provided a very useful taxonomy for characterising different group tasks in terms of the way in which individual performances are translated into the group outcome. Steiner draws a distinction between compensatory and non-compensatory task environments. In compensatory tasks, the performance of a weaker team member can be compensated by a stronger member in the team. In non-compensatory tasks, in contrast, the group performance is determined by the weakest member. Compensatory tasks can be further distinguished by whether the performance of all members in the team affects the team.

Tug-of-war is a prototypical example for a compensatory task in which the performance of all team members matters. Each players' individual performances combine in an additive fashion to determine the team's overall strength. The weakness of one member can be compensated by the strength of another. A pub quiz is an example for a task in which performance is compensatory but the group outcome does not depend on each member. In order for a group to do well, only one person needs to know the correct answer on each round. Hence, individual answers combine in a disjunctive fashion for positive outcomes (i.e. only one correct answer is necessary) and in a conjunctive fashion for negative outcomes (i.e. the team only fails if none of the players knows the correct answer). Finally, there are non-compensatory group tasks, such as a climbing team, for which the group performance depends on everyone in the group and is determined by its weakest member. Here, the group performance function is conjunctive for a positive outcome and disjunctive for a negative outcome.

The rest of this section is organised as follows. We will first describe different candidate models of responsibility attribution in groups that make predictions about participants' attributions in group contexts such as the situations described above.[6] We will then describe an experiment in which we investigated how different causal structures influence how responsibility is distributed between multiple agents. We will analyse the results in multiple ways to shed light into what systematic aspects of the data the models can capture and what they miss out on. The discussion of these findings will motivate a series of follow-up experiments which will be described in later sections of

[6]The reader may have noticed the switch from 'I' to 'We'. Generally, I will use 'We' when describing empirical work in which I have benefitted from the collaboration with others. I will use 'I' for personal speculations and comments for which my collaborators deserve no blame.

this chapter.

4.1.4 Modelling responsibility allocations between multiple agents

There are multiple ways in which one could try to explain how people attribute responsibility in group settings such as the ones described above. Here we will focus on three models which differ in terms of (i) how they identify individual agents as causes of the group outcome and (ii) how they distribute responsibility between the identified causes.

4.1.4.1 Matching model

The Matching model identifies each agent that is associated with the outcome as causal (see Heider, 1958 and Shaver, 1985 who both regard *association* as the lowest possible level of causal connectedness). The Matching model then attributes responsibility to each identified cause as a direct linear function of that agent's performance. That is, the better an agent performed the more responsibility she receives for a win and the less responsibility for a loss. The matching model assumes no knowledge of the underlying causal structure.

4.1.4.2 Counterfactual model

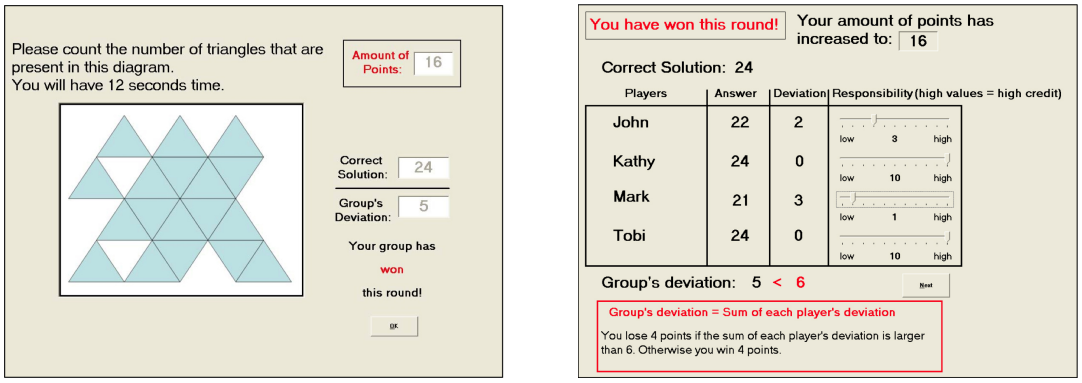
The Counterfactual model is based on the intuition that in order to be responsible for an outcome, an agent's contribution needs to have made a difference to the outcome. Hence, it uses the counterfactual *but for* test (Hart & Honoré, 1959/1985; Lewis, 1973) on each agent to see whether the team outcome would have been different, had the agent of interest acted differently. If the answer to the *but for* test is positive then the agent is seen as a cause, otherwise not. Each identified cause is then consequently assigned full responsibility.

4.1.4.3 Structural model

The Structural model (Chockler & Halpern, 2004) was explicitly developed to capture people's intuitions of how responsibility should be attributed in situations of overdetermination. As we have seen in Chapter 3, the model operates via first modelling the situation in terms of a causal network which captures the causal relationships between agents and outcome. It then uses a relaxed criterion of counterfactual dependence to identify the actual causes of the outcome. According to this criterion, an event C counts as a cause for an event E not only if there is a counterfactual dependence between E and C in the *actual situation* but also if there would have been a counterfactual dependence between E and C in a *possible situation* in which the values of other variables in the causal structure (apart from C and E) were fixed at a certain value. The degree to which C is responsible for E is a function of the number of changes that are necessary

to transform the actual world (in which there is not counterfactual dependence between E and C) into a possible world in which E counterfactually depends on C (i.e. a world in which C is pivotal for E). More specifically, the responsibility of C for E is given by: Responsibility(C,E) $= \frac{1}{N+1}$, whereby N equals the minimal number of changes that are necessary to make C pivotal for E.

4.1.5 Experiment



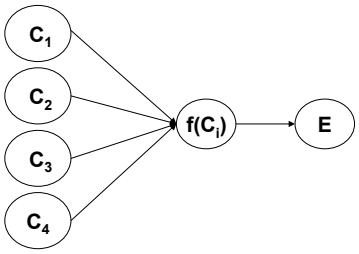
(a) Game Phase **(b)** Attribution Phase

Figure 4.1: Screenshots of the two phases of the Triangle Game. In the Game Phase, participants have to count how many triangles are present in the stimulus. In the Attribution Phase, participants assign responsibility to each player for the team's outcome.

To investigate how people attribute responsibility in group contexts we developed the Triangle Game. Four players form a team and each player faces the same task, namely to count the number of triangles in a complex geometrical stimulus such as the one displayed in Figure 4.1a within a limited amount of time. Note that not only the small blue triangles count but also lager blue triangles that are made out of smaller ones. The diagram in Figure 4.1a, for example, contains 24 triangles (18 small blue triangles, 5 middle sized ones, and 1 big one). Each player's score is given by the error of their estimate from the true number of triangles that the diagram contained. Hence, the lower a person's error score the better.

Figure 4.2 shows a causal graph that represents the underlying structure of the Triangle Game. There are four players in the team (C_1 to C_4) who independently contribute to the team outcome (E). There are different ways in which the individual causes can combine to bring about the effect.[7] According to the *sum* function, the group outcome is positive as long as the sum of each person's error is less than 7 (see Table 4.1, first row). An individual player's error is defined as the absolute difference between their estimate and the correct answer. The *sum* function creates a compensatory task

[7]Note that the integration function $f(C_i)$ would normally not be explicitly represented in the graph structure but merely captured in the structural equations that are associated with the graph.



f(Ci)	Win	Loss
sum(Ci)	additive	additive
min(Ci)	disjunctive	conjunctive
max(Ci)	conjunctive	disjunctive

Figure 4.2: Different integration functions (summation, minimum, maximum) determining how multiple causes (C_1 to C_4) combine to bring about an effect (E = {win, loss}).

environment. If one person has a high error this can be compensated by low errors of the other team members.[8] The individual player's scores combine in an additive fashion both for team wins and losses. According to the *min* function, the team wins if at least one player gave the correct solution (see Table 4.1, second row). The *min* function also establishes a compensatory task environment. However, in contrast to the *sum* function in which the outcome always depends on the performance of each player in the team in an additive fashion, the *min* function renders some players' performances potentially superfluous.

Finally, the *max* function creates a task environment that is non-compensatory and in which the outcome depends on the performance of each player (see Table 4.1, third row). The team wins only if all players manage to perform quite well. If one ore more player's error is larger than 2, the team loses.

Table 4.1: Some core properties of different combination functions.

function	compensatory	all players required	mathematical form
sum	yes	yes	$E = \begin{cases} 1 & \text{if } sum(C_i) < 7 \\ 0 & \text{otherwise} \end{cases}$
min	yes	no	$E = \begin{cases} 1 & \text{if } min(C_i) = 0 \\ 0 & \text{otherwise} \end{cases}$
max	no	yes	$E = \begin{cases} 1 & \text{if } max(C_i) < 3 \\ 0 & \text{otherwise} \end{cases}$

Let us consider the situation depicted in Figure 4.1 to illustrate the different combination functions. The correct solution for the depicted stimulus is 24 triangles. John's

[8] Unless, of course, an individual player happens to make the team lose through a very bad performance (i.e. an error of 7 or more).

answer was 22 and hence he has an error of 2. Kathy gave a correct estimate and hence has an error of 0. Mark has an error of 3 and Tobi and error of 0. In the *sum* condition, the group would have won this round since $sum(2, 0, 3, 0) < 7$. The team would have also won this round in the *min* condition since $min(2, 0, 3, 0) = 0$. However, the team would have lost this round in the *max* condition because $max(2, 0, 3, 0) \geq 3$.

In the experiment, each round of the game consisted of two consecutive steps. In the first step, participants were shown the diagram for $12s$ (see Figure 4.1a). They had to count the triangles and type in their answer. They then saw the correct solution and were informed about whether their group won or lost the round. The team's points tally changed accordingly. For each round that the team won, they gained a number of points and for each round that team lost, the team's number of points was reduced. Participants were instructed that the aim of the game was to win as many points as possible. In the second step, participants viewed a table, which contained the answers of each player in their group (see Figure 4.1b). Furthermore, participants could see how much each player's answer deviated from the correct solution. Participants used sliders to indicate each player's responsibility for the group's loss or win in that particular round (including themselves). The scale ranged from 0 ('not responsible at all') to 10 ('very responsible'). The participant's score was always listed at the bottom of the table.

The Triangle Game allows us to investigate the influence of individual performance and differences in causal structure while controlling for many other factors. Furthermore, the game provides a challenging test ground for models of responsibility attribution: multiple individual causes combine in several different ways and the group outcomes are often overdetermined (e.g. when two or more players give the correct solution in the *min function* condition).

4.1.5.1 Methods

Participants and materials Sixty-nine participants were recruited via E-Mail and played the Triangle Game on individual computers. They participated for the chance of winning one of three prizes in a £150 draw. The experiment was programmed in Visual Basic 6.0.

Procedure In a between-subjects design, participants were randomly assigned to one of the three experimental conditions (*sum*: $N = 21$, *max*: $N = 23$ or *min*: $N = 25$), which only differed in terms of the combination function. Participants were instructed about the combination function and were able to remind themselves throughout the game (see Figure 4.1b bottom). Each round of the game consisted of two steps: first, the triangle count and then the responsibility attribution as described above. The game finished after 11 rounds were played including an initial practice round.

4.1.5.2 Results

We will first look at participants' responsibility attributions as a function of a player's performance and the different combination functions separately for wins and losses. We will then look at how participants attributed responsibility in situations of overdetermination specifically before testing the predictions of the responsibility models introduced above.

Mean responsibility attributions Figure 4.3a shows the mean responsibility rating assigned to a player for different error scores in rounds that the group lost. Figure 4.3b shows these ratings for rounds that the group won. To establish whether participants gave different patterns of responsibility ratings in the three conditions we conducted separate mixed ANOVAs for losses and wins, with Condition (sum, max, min) as between-subject factor and Error as within-subject factor. Comparisons were only made for values of Error in which responsibility ratings were available for all three conditions, that is for an error of $1-4$ for losses and $0-2$ for wins.

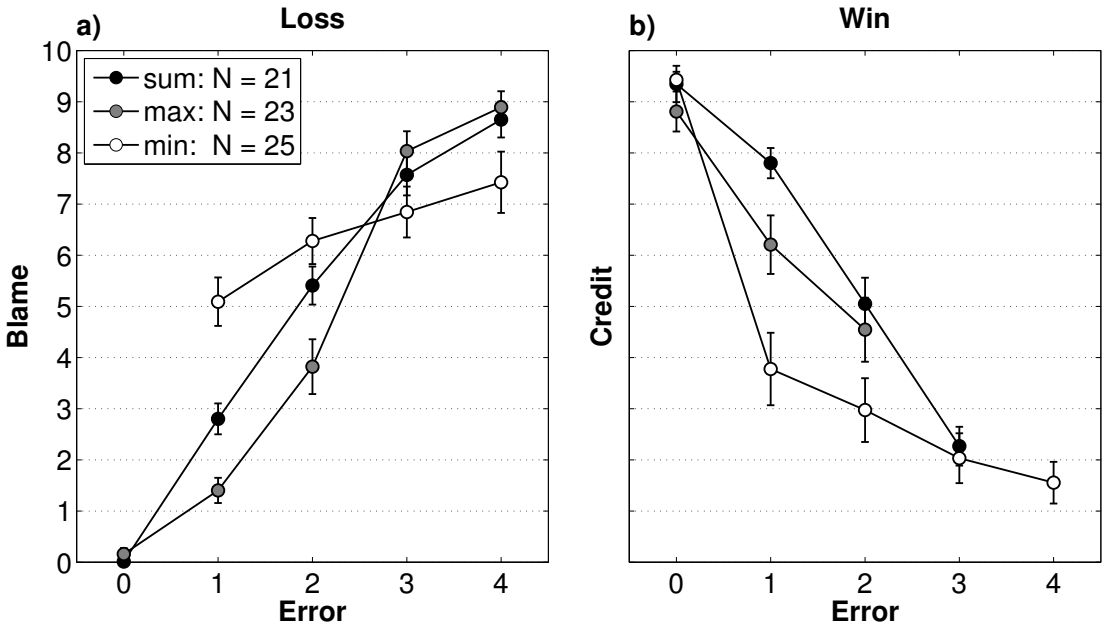


Figure 4.3: Mean responsibility attributions separated for (a) losses (blame) and (b) wins (credit) as a function of each player's error (x-axis) and the different combination functions. Error bars are ± 1 *SEM*.

For losses there was no effect of Condition, $F(2,66) = 1.92, p = .154, {\eta_p}^2 = .055$, but an effect of Error, $F(3,198) = 152.65, p < .001, {\eta_p}^2 = .698$, and an interaction between Condition and Error, $F(6,198) = 16.67, p < .001, {\eta_p}^2 = .336$. For wins there was a significant effect of both Condition, $F(2,66) = 6.27, p = .003, {\eta_p}^2 = .160$ and Error, $F(2,132) = 129.15, p < .001, {\eta_p}^2 = .662$ as well as an interaction effect between Condition and Error, $F(4,132) = 8.07, p < .001, {\eta_p}^2 = .196$.

Figure 4.3a shows that the effect of error for losses is due to the general trend of attributing more responsibility for increased error values, which held in all three conditions. Indeed, the linear contrast for Error was significant, $F(1, 66) = 250.85, p < .001, {\eta_p}^2 = .792$, as well as the linear contrast for the interaction between Error and Condition, $F(2, 66) = 25.44, p < .001, {\eta_p}^2 = .435$. Conversely, for wins (see Figure 4.3b) responsibility attributions generally decreased with an increased error. Again, the linear contrasts for both Error, $F(1, 66) = 190, p < .001, {\eta_p}^2 = .742$, and the interaction between Error and Condition, $F(1, 66) = 4.16, p = .02, {\eta_p}^2 = .112$, were significant. The significant interaction for both wins and losses shows that the relationship between responsibility ratings and a player's error were qualitatively different between the three conditions.

For losses, responsibility attributions increased in a clear linear fashion in the *sum* condition. In the *max* condition the same general linear trend held. However, there was a greater difference between how much responsibility a player received for an error of 2 ($M = 3.82, SD = 2.56$) and 3 ($M = 8.03, SD = 1.86$) compared to the *sum* condition ($M = 5.41, SD = 7.57$ for an error of 2 and $M = 7.57, SD = 1.83$ for an error of 3). Finally, in the *min* condition, participants attributed a high degree of responsibility for the loss even for a small error of 1 ($M = 5.09, SD = 2.36$) and the steepness with which attributions increased as a function of error was smaller than in the *sum* and *max* conditions.

For wins, responsibility ratings in both the *sum* and *max* conditions decreased linearly with an increasing error. In the *min* condition, however, only players who had an error of 0 received high responsibility ($M = 9.43, SD = 0.79$) and attributions dropped for an error of 1 ($M = 3.77, SD = 3.54$) or more.

Overall, these analyses establish that the experimental variation had an influence on the general trend of responsibility attributions. However, they cannot reveal how people distribute responsibility given specific configurations of players' answers. We will now look at these cases in more detail.

Situations of overdetermination Given the setup of the Triangle Game, situations of overdetermination occur naturally. As outlined above, the Structural model (Chockler & Halpern, 2004) predicts that a player's responsibility is reduced in situations in which the outcome is overdetermined. More specifically, the model predicts a clear trend in that an individual's responsibility decreases the more changes would be necessary to render her pivotal.

Table 4.2 shows the predictions of the different models outlined above for situations of overdetermination in the *min* and *max* condition. Since the Matching model is insensitive to the causal structure, it attributes equal responsibility to the same score independent of whether or not the outcome was overdetermined. The Counterfactual model, in contrast, only assigns responsibility to a player if she was pivotal. As soon as

Table 4.2: Predictions by the different models for situations of overdetermination.

	Win in *min* condition			**Loss in *max* condition**		
	$\#1 = 0$	$\#2 = 0$	$\#3 = 0$	$\#1 > 2$	$\#2 > 2$	$\#3 > 2$
Matching	1	1	1	1	1	1
Counterfactual	1	0	0	1	0	0
Structural	1	1/2	1/3	1	1/2	1/3

Note: # = number of players who had a certain error.

an outcome is overdetermined, none of the players receive any responsibility. Finally, the Structural model assigns responsibility as a function of the minimal number of changes that are necessary to render a player pivotal. Hence, responsibility reduces the more the outcome is overdetermined.

For example, in the *min* condition (in which the team wins if at least one player's error equals zero), the team's win is overdetermined if more than one player gives the correct answer. In line with the predictions of the Structural model, a player received most responsibility in situations in which she was the only one who gave the correct solution ($M = 9.89, SD = 0.29$) compared to situations in which two players ($M = 8.89, SD = 1.42$) or three players ($M = 7.96, SD = 1.74$) gave the correct answer, $F(2, 48) = 17.92, p < .001, {\eta_p}^2 = .428$ (see Figure 4.4 light grey bars).

Although the effect is smaller than predicted by the Structural model, the linear contrast is significant, $F(1, 24) = 30.83, p < .001, {\eta_p}^2 = .562$. Similarly, in the *max* condition, the team's loss is overdetermined if more than one player gives an answer with an error of more than two. Again, as predicted by the Structural model, a player received most responsibility in situations in which he was the only one to have an error

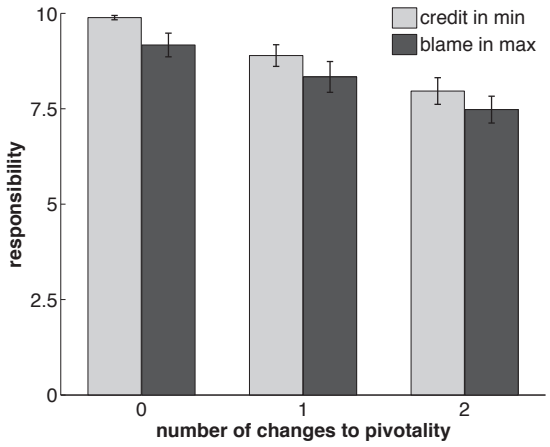


Figure 4.4: Mean responsibility attributions to individuals for wins in the *min* condition (grey bars) and losses in the *max* condition (black bars) as a function of the number of changes to pivotality. Error bars indicate ±1 *SEM*.

greater than two ($M = 9.17, SD = 1.49$) compared to situations in which there were two players ($M = 8.33, SD = 1.93$) or three players ($M = 7.48, SD = 1.69$), $F(2, 44) = 14.71, p < .001, {\eta_p}^2 = .401$ (see Figure 4.4 dark grey bars). Again, a significant linear contrast demonstrates that responsibility attributions to a player decreased the more changes were necessary to render him pivotal, $F(1, 22) = 25.22, p < .001, {\eta_p}^2 = .534$.

These results show that participants' responsibility attribution are generally in line with the predictions of the Structural model for situations of overdetermination.[9] The Counterfactual model cannot predict this trend as it assumes that players only receive responsibility when the outcome is not overdetermined. However, of course there are many possible situations in the game in which the outcome is *not* overdetermined. We will now look at how well the different models explain participants' attributions overall.

Model predictions Table 4.3 and Table 4.4 show the predictions of all three models for one particular situation in which a team lost and won, respectively. The model predictions were scaled to fit the range of participants' response scale.

Table 4.3: Model predictions for a loss situation for the three experimental conditions.

		sum			***max***			***min***		
Player	Error	M	C	S	M	C	S	M	C	S
John	3	7	0	5	7	0	10	7	10	10
Kathy	4	6	10	10	6	0	10	6	10	10
Mark	1	9	0	5	9	0	0	9	10	10
Tobi	2	8	0	5	8	0	0	8	10	10

M = Matching model, C = Counterfactual model, S = Structural model.

As noted above, the Matching model (MM) makes the same predictions independent of the combination function. It predicts that responsibility attributions for losses increase and for wins decrease the greater the error. The Counterfactual model (CM) predicts that a player is assigned full responsibility if she is pivotal and no responsibility otherwise. The Structural model (SM) assigns full responsibility to pivotal players and reduced responsibility to non-pivotal players. We will now discuss the predictions of the CM and SM for the different combination functions separately.

For losses in the *sum* condition, only pivotal players are assigned responsibility by the CM. However, the SM also assigns (reduced) responsibility to players who could have made the team win, had the performance of the other players been better. For wins, both the CM and the SM predict that each player will receive full responsibility

[9] Note, however, that the Structural model predicts a logistic rather than a linear decrease of responsibility as a function of the extent to which the outcome is overdetermined. The data only support the general prediction that responsibility decreases. They do not support the prediction about the exact functional form.

Table 4.4: Model predictions for a win situation for the three experimental conditions.

		sum			***max***			***min***		
Player	Error	M	C	S	M	C	S	M	C	S
John	0	10	10	10	10	10	10	10	0	10
Kathy	2	8	10	10	8	10	10	8	0	0
Mark	0	10	10	10	10	10	10	10	0	10
Tobi	2	8	10	10	8	10	10	8	0	0

M = Matching model, C = Counterfactual model, S = Structural model.

(independent of a player's error). Each player is pivotal in such a situation and could have made the team lose had their answer been worse.

For losses in the *max* condition, the CM and the SM predict that a player will receive no responsibility if their error was less than three. If only one player's error was greater than two both models predict that this player will receive full responsibility. However, if two or more players erred more than two, the SM predicts that each player's responsibility reduces whereas the CM predicts that the players will receive *no* responsibility because the loss is overdetermined. For wins, both models predict that each player will receive full responsibility. Again, if the team won, all players are pivotal – the team would have lost if any of the players' error had been greater than two.

In the *min* condition, the CM and the SM predict that each player receives full responsibility for losses. Each of the players is pivotal in such a situation and could have made the team win by giving the correct answer. For wins, both models predict that players who did not give the correct answer will receive no responsibility at all. If just one player gave the correct answer both models predict that this player will receive full responsibility. If more than one player gave the correct answer, the SM predicts that the responsibility will be reduced whereas the CM predicts that none of the players will receive any responsibility because the win is overdetermined.

Table 4.5 shows the correlation of each model with participants' ratings separately for the three experimental conditions.[10] The Structural model predicts participants' ratings better than the Counterfactual model in each experimental condition. However, participants' ratings are even better explained by the Matching model.

An inspection of Figure 4.3 reveals why the MM fares better than the CM and SM. While there are significant differences between the experimental conditions, the general trends are similar between conditions. Responsibility attributions for losses increase with an increased error and attributions for wins decrease with an increased error as predicted by the MM for all three experimental conditions. As predicted by the MM, responsibility attributions changed linearly with a player's performance in the

[10]In order to correlate model predictions with participants' ratings, we z-scored the data on the level of individual participants and the predictions of each model.

Table 4.5: Model correlations for the three different experimental conditions.

	sum	***max***	***min***
Matching (MM)	0.84 (0.53)	0.8 (0.60)	0.51 (0.94)
Counterfactual (CM)	0.64 (0.81)	0.48 (0.97)	0.27 (1.14)
Structural (SM)	0.68 (0.76)	0.61 (0.83)	0.38 (1.06)

Note: Values in parentheses are root mean squared errors.

sum condition.

Similarly, in the *max* condition, responsibility attributions for losses increased with greater errors. However, there was a greater step between an error of two and three compared to the *sum* condition. In contrast to the predictions by the CM and the SM, players whose answers were in the range between zero and three still received some responsibility for the team's loss. While both the CM and SM predict that each player will receive full responsibility when the team won, attributions decreased with an increased error as predicted by the MM.

Finally, for losses in the *min* condition, there was also a small trend for an increase in responsibility for greater errors. Both the CM and SM predict that each player will be held fully responsible independent of their error. For wins, the CM and SM predict that only a player who gave the correct answer will be held responsible. However, participants also attributed some responsibility to players who did not give the correct answer whereby attributions decreased with an increased error.

4.1.5.3 Discussion

This study established the first empirical test of the Structural model of responsibility attribution (Chockler & Halpern, 2004). The results show that participants' responsibility attributions to individuals in groups are both influenced by an individual's performance as well as the way in which performances within the group combined to determine the outcome. Players are attributed different degrees of responsibility for the same performance depending on the combination function. The Structural model captures many of the qualitative trends in the data such as the reduction of responsibility in situations of overdetermination. However, it cannot capture the fact that responsibility attributions vary as a function of performance even in situations when all players are pivotal. For example, in the *max* condition, responsibility attributions for wins decrease with an increased error from $M = 8.81$ ($SD = 1.88$) for an error of zero to $M = 4.54$ ($SD = 3$) for an error of two. Hence, while any player could have made the team lose in these situations and is thus pivotal, participants still differentiated between players with different scores.

While this result is clearly at odds with the Structural model, it is easy to see how

a more nuanced counterfactual account would be able to capture this trend. Intuitively, a player with an error of two was closer to making the team actually lose the game compared to a player who gave the correct answer. The counterfactual in which a player moves from an error of zero to three is clearly more distant from the actual situation compared to a situation in which a player's error changes from two to three. This result suggests that responsibility attributions are not only sensitive to how difficult it is to make a player pivotal via changing other variables but also to how difficult it is to change the player under consideration in order to change the outcome. More specifically, responsibility attributions appear to increase the less it takes to make a player pivotal and decrease the less it takes to change that player's answer in order to flip the group outcome. The same sort of reasoning can be applied to account for the fact that players who did not cause the actual outcome still received some responsibility. While it is true that they did not cause the outcome in the actual situation it is easy to imagine how they could have. For example, a player who had an error of one in the *min* condition was very close to helping to make the win.

It is worth pointing out some limitations of the experiment. The setup of the game was such that performance was measured in terms of a discrete variable expressing the difference of each person's answer from the correct solution. While this setup allowed us to directly compare the *sum*, *min* and *max* condition, it generated a difficult test environment for the Structural and Counterfactual model. As outlined in Chapter 3, the Structural model was originally developed to deal with binary variables only. Indeed, counterfactual models in general have problems with situations in which individual variables can take on more than two values as *the* counterfactual state for a variable is not well-defined anymore.

As mentioned above, this problem is particularly apparent in the *sum* condition when the team succeeded. Both models predict that each player is fully responsible because they could have made the team lose had their performance been worse.[11] However, if the team succeeded, the model does not distinguish between a player with an error of zero and a player with an error of 6. The same is true for a loss in the *min* condition. Again, since all players are pivotal in such a situation they are predicted to be fully responsible by both counterfactual models no matter whether their error was one or a hundred. Further theoretical developments are required before counterfactual accounts can become viable accounts for attributions of responsibility in continuous worlds. For this reason, we will resort to situations which can be modelled with binary variables in subsequent experiments in this chapter.

[11] When we use the term 'fully responsible' we do not mean that a person is fully responsible in the sense that she subsumes all the responsibility for the group outcome (and there is no responsibility left for the other group members). Rather, we mean with 'fully' that a person is responsible without any diminution. Thus, when a person in a group is fully responsible for an outcome, we mean that this person is as responsible as he or she would be if the outcome had just been determined by herself without anyone else being present (see Zimmerman, 1985).

Having discrete rather than binary variables is also problematic for the notion of a *minimal change* between two worlds on which the Structural model relies. Recall that the minimal change to render a person under consideration pivotal is defined in terms of the smallest number of variables whose value needs to be changed. However, as discussed in Chapter 3, this notion becomes problematic when variables are not binary. For example, imagine that the error scores of the team members in the *max* condition are 0, 7, 2 and 3. In order to make the last player pivotal for the loss, we need to change one variable. Contrast this with the situation in which the error scores are 0, 3, 3 and 3. In this situation, *two* variables need to be changed in order to make the last player pivotal. However, arguably, the change in the latter situation is *smaller*.

In a follow-up experiment, we directly addressed the question of how participants conceptualise the notion of a minimal change between situations (Gerstenberg, 2009). Participants again played the triangle game and attributed responsibility afterwards. However, we added a third stage in which participants were asked to make as few changes a possible to the error scores of one or more players so that the outcome in this round would have been different. That is, by making adjustments to the error scores of one or more players, participants were asked to generate a loss when the original result was a win or vice versa. We found that in situations in which participants could bring about the alternative outcome in different ways, they preferred to make smaller changes to more variables rather than larger changes to fewer variables. More precisely, when having to change the outcome from a loss to a win, participants preferred to make small changes to several variables rather than one big change to a single variable in 94.4% of the times. Similarly, when the result needed to be changed from a win to a loss, 84.6% preferred to make several small changes rather than one big change.[12]

Finally, because we wanted the experiment to be engaging, we had participants form part of the team. However, this created a problem in that participants now partly determined the outcome of the game. For example, a participant who was particularly good at counting triangles would experience more wins than a participant who found the task more difficult. Furthermore, the extent to which participants were in control of the outcome differed between conditions. We had fixed the answers of the computer players in a way that would yield high chances of having an equal number of wins and losses in the three conditions (for a range of possible responses by the participants). Nevertheless, participants had the chance to always make the team win in the *min* condition as a single correct answer is sufficient for the team win. In contrast, in the *max* condition, some of the team losses were already determined if at least one of the other players had an error of three or more. In these situations, participants had no control over the outcome. In fact, whereas only 36% of the rounds in the *max* condition were won, the team was successful in 63% of the rounds in the *min* condition – compared

[12]For this analysis, we selected only cases in which changing one variable would have been sufficient to reverse the outcome and in which at least two error points needed to be changed.

to 46% in the *sum* condition. Hence, while participants in the *min* condition might have thought that the team task is quite easy, participants in the *max* condition might have found it quite hard. Moreover, differences in participants' performance will have affected their perceived skill level of the other players in the group. A participant who found the task easy might have been inclined to give high responsibility for a loss to the same person to which another participant, who found the task hard might have given only minimal responsibility.[13] In order to increase experimental control and facilitate comparability between the attributions of different participants, participants acted as external observers in all subsequent experiments in this chapter.

4.2 Responsibility Attributions in Asymmetric Structures (Zultan et al., 2012)

The results of the previous experiment have provided some support for the use of causal reasoning in responsibility attribution.[14] People's attributions were systematically affected by the way in which individual contributions combined to yield the group outcome. Furthermore, the structural model of responsibility (Chockler & Halpern, 2004) predicted participants' attributions better than a simple counterfactual model.

However, since the roles of each player within the different conditions were symmetrical, some of the more subtle predictions by the structural model could not be tested. Furthermore, some of the trends found in the data (such as the reduction of responsibility in situations of overdetermination) could also be explained by alternative, non-causal accounts. For example, a *diffusion of responsibility* account (Darley & Latané, 1968), which just assumes that responsibility to each individual in the group reduces the more it is shared with other people would also yield this prediction. For a simple diffusion of responsibility account, the structure of the situation and the relationships between the individuals in the group does not matter. Generally, the more people failed their task the less each person in the group is predicted to be held responsible for the negative outcome. In contrast, the structural model predicts that the causal dependencies between the contributions of each individual in the group are crucial for how responsibility is attributed. As we will see, the model predicts that a person will sometimes be held *less* responsible for a negative group outcome even if *fewer* people caused the negative outcome. A simple diffusion of responsibility account cannot make this prediction – for such an account, responsibility *increases* the fewer people caused the loss.

The experiments in this section use asymmetric task structures to ascertain the

[13] Note that participants' performance did not vary as a function of the experimental condition. Participants' average error was $M = 1.5, SD = 1.67$ in the *sum*, $M = 1.47, SD = 1.41$ in the *max* and $M = 1.61, SD = 1.46$ in the *min* condition, $F(2, 66) = .479, p = .622, {\eta_p}^2 = .014$.

[14] This section has greatly benefited from Ro'i Zultan's contribution who is the first author of the corresponding article. Parts of this section are reprinted from Zultan et al. (2012) and have been written in collaboration with Ro'i Zultan and David Lagnado.

influence of causal reasoning for attributions of responsibility and to rule out non-causal explanations such as a simple diffusion of responsibility. In a recent study, Forsyth et al. (2002) have investigated how responsibility diffuses in cooperative collectives. Generally, their results supported the diffusion of responsibility idea in that responsibility decreased for larger groups. Note, however, that this result was implied by the methodology as participants were asked to distribute a fixed number of 100 points between the members of the group. Naturally, individual group members receive more points in smaller groups than in larger groups. For this reason, we do not fix the overall amount of responsibility that is to be assigned in our experiments. Hence, participants can in principle attribute a high (or low) degree of responsibility to each member of the group. More interestingly, Forsyth et al. (2002) found that responsibility diffused unequally between the members of the group. However, they were unable to identify any factors that explained the observed patterns of responsibility allocations. They conclude that "further research must identify the relationships between the type of task the group completes – and in particular, the way the task constrains the way group members' individual contributions are combined – and responsibility allocations" (Forsyth et al., 2002, p. 65). This is exactly what the following experiments are set out to do.

4.2.1 Theoretical framework

For the reasons outlined above, we will restrict our attention to situations in which individual performances as well as the group outcome can be represented in terms of binary variables. Hence, individuals can either *fail* or *succeed* in their tasks and the team as a whole can either *win* or *lose* the challenge. Again, we will look at cases in which there are no direct causal dependencies between the players in the group. That is, the performance of one player does not directly affect the performance of another player.

Let us denote the outcome o of an individual player i by $o_i \in \{0, 1\}$, with $0 =$ *failure* and $1 =$ *success*. The team outcome t is determined by a team function $t = f(o_1, o_2, \ldots, o_n) \in \{0, 1\}$, with $0 =$ *loss* and $1 =$ *win*. The function f is weakly increasing in o_i, that is, the team outcome cannot benefit from a *failure* of a team member, and similarly cannot be harmed by any of the team members *succeeding*.

We will see in the following that this basic framework is rich enough to capture the principles of simple causality, counterfactual causality, and diffusion of responsibility. We will consider several models that differ in how they take into account peer performance and causal relationships when assigning responsibility to any one team member.[15]

[15]The experiments in this section will focus on the attribution of blame for negative outcomes. As discussed in Chapter 3, Chockler and Halpern (2004) distinguish between *responsibility* and *blame*. Blame is defined as anticipated responsibility and thus relative to the epistemic state of the agent. However, since agents' epistemic states are not varied, blame will be equated with negative responsibility.

4.2.1.1 Simple responsibility (SimResp)

The benchmark model ignores both peer performance and the causal structure. The model simply assigns a responsibility of 1 if the individual and team outcome are aligned, and 0 otherwise. In other words, if the team lost, then all the team members who have failed their individual task receive blame, and if the team won, all the team members who have succeeded receive credit. It can be seen as a binary version of the Matching model in the Triangle Game experiment (Gerstenberg & Lagnado, 2010).

4.2.1.2 Diffusion of responsibility (DiffResp)

The diffusion of responsibility model also ignores the causal relationships but takes into account peer performance as it divides the responsibility equally between all individuals who are assigned full responsibility by SimResp. The model can be interpreted as a normalised version of SimResp, in which the total responsibility sums up to exactly 1.

4.2.1.3 Simple pivotality (SimPiv)

The simple pivotality model refines SimResp by imposing a further condition on responsibility, namely, the model assigns a responsibility of 1 if and only if the individual and team outcomes are aligned *and* the individual is pivotal. Remember that an individual is pivotal if there is a counterfactual dependence between the outcome of the team and the individual. Hence, the SimPiv model assigns blame only to team members who failed but could have made their team win had they succeeded, given the performance of their peers. The simple pivotality is equivalent to Gerstenberg and Lagnado's (2010) Counterfactual model.

4.2.1.4 Counterfactual pivotality (CFPiv)

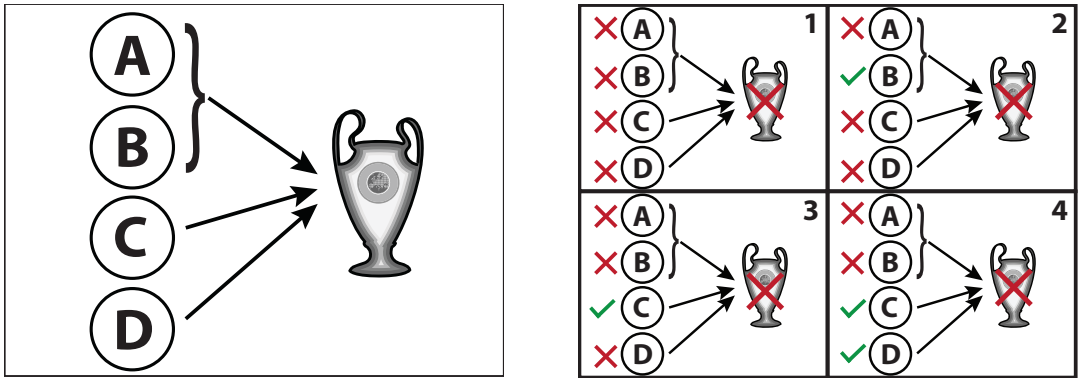
The counterfactual pivotality model is equivalent to Gerstenberg and Lagnado's (2010) Structural model, which is derived from Chockler and Halpern's (2004) general model of responsibility. Similar to SimPiv, CFPiv assigns a responsibility of 1 to individuals who are pivotal. The two models differ with regard to individuals who are not pivotal, but whose outcome is aligned with their team's outcome. Whereas non-pivotal individuals receive 0 responsibility according to SimPiv, responsibility reduces with the minimal number of changes to pivotality according to CFPiv. Hence, the responsibility of an individual player's outcome o_i for the team outcome t is given by: $Resp(o_i, t) = \frac{1}{N+1}$, whereby N equals the minimal number of changes to pivotality.

4.2.2 Experiment 1

Apart from the SimResp model which serves as a benchmark, all other models take peer-performance into account when attributing responsibility. However, they do so in

different ways. Whereas the SimPiv and the CFPiv model are sensitive to the causal structure of the situation, the DiffResp model is only sensitive to the number of players who failed in their tasks. Nevertheless, as we have seen above, the CFPiv and DiffResp model can yield identical predictions for situations in which the causal structure is symmetric. For example, consider a situation in which all four members need to succeed in order for the team to win the challenge (i.e. $t = min(o_1, o_2, o_3, o_4)$). If all individuals failed, the DiffResp model assigns blame of $\frac{1}{4}$ to each individual. The CFPiv model makes the same prediction since $N = 3$ changes are necessary to render each player pivotal.

In order to clearly dissociate the predictions of the different models, we will look at situations in which the individual players' roles in the team are asymmetric. Whereas for some players their individual success is necessary for the team to win, for other players, the team still has a chance of winning even if they did not succeed in their tasks. The CFPiv model predicts that the effect that one player's performance has on another player's responsibility for the outcome depends on the exact relationship between the two players. Assuming that the team has lost the challenge, if two players are *substitutes*, such that the success of one of them makes the success of the other unnecessary for the team winning, each team member must fail in order for the other to be pivotal, $t = max(o_1, o_2)$. Therefore the success of one reduces the blame assigned to the other. Conversely, if the two team members are *complementary*, so that in order for one to be pivotal the other must succeed, the success of one increases the blame assigned to the other, $t = min(o_1, o_2)$.



(a) Team challenge **(b)** Different situations

Figure 4.5: (a) Asymmetric team challenge used in Experiments 1 and 2. For the team to win, both C and D have to succeed and at least one out of A and B (i.e., $t = min(max(o_A, o_B), o_C, o_D)$); (b) Four different situations in which the team has lost the challenge. *Note*: ✗= failure, ✓= success.

Consider the team challenge shown in Figure 4.5a. In this challenge, both player C and D and at least one out of A and B have to succeed in their individual tasks in order for the team to win the challenge (i.e. $t = min(max(o_A, o_B), o_C, o_D)$). Figure

4.5b shows four different situations in which the team has lost the challenge. Because A failed in each situation, SimResp predicts that A will be blamed. In contrast, SimPiv predicts that A will only be blamed in the fourth situation, the only situation in which A is pivotal. According to the DiffResp model, A's blame in situation 1 is $\frac{1}{4}$ because there are four players who failed their tasks and amongst whom the blame for the loss is shared. A's blame is predicted to be $\frac{1}{3}$ in situations 2 and 3 and $\frac{1}{2}$ in situation 4.

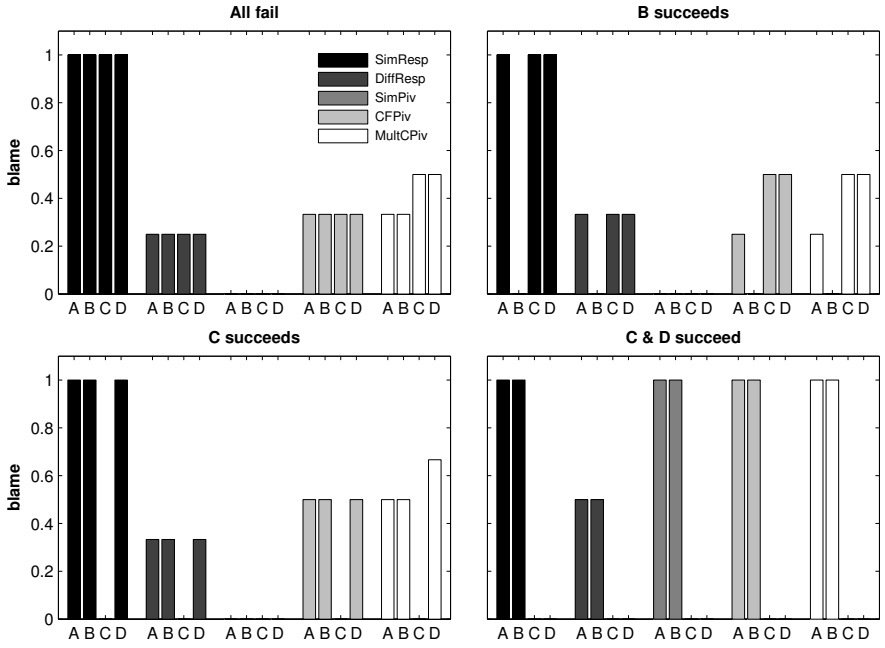


Figure 4.6: Predicted blame for players A, B, C and D by the different models for the four different situations shown in Figure 4.5b.

According to the CFPiv model, A is predicted to receive a blame of $\frac{1}{3}$ in situation 1. A minimum of two changes are required to render A pivotal in this situation (i.e. changing C and D). In situation 2, A's blame is predicted to *decrease* although less players failed in their task (in contrast to DiffResp which predicts an *increase* in blame). An additional change is required, compared to situation 1, in order to make A pivotal. All other players need to be changed (including a change of B to having failed) to generate the only situation in which A is pivotal for the loss. Hence, A's blame is predicted to be $\frac{1}{4}$ in this situation. In situation 3, in which C succeeded in her task, A's blame is predicted to increase compared to situation 1. Now, only one change is needed to render A pivotal and A's blame is predicted to be $\frac{1}{2}$. Finally, in situation 4, A is pivotal and his blame thus predicted to be 1. Figure 4.6 summarises the model predictions of all the models for the four different situations.[16] As the figure shows, the different models not only make different predictions about how the blame attributed to A will change between the different situations but also about how blame will be apportioned between

[16]The MultCFPiv model will be discussed below.

the players within a situation. We will come back to this point later.

Table 4.6: Scenario and questions for Experiment 1. The labels "all fail", "B succeeds" and "C succeeds" were not part of the original materials.

In a new cooking show on television, a group of four chefs are charged with the task of preparing a meal in a certain culinary style. A meal is composed of two starters, one main dish and a dessert. The show panel judges each of the four dishes, and determines whether it's successful or not. The group wins the task if the meal is successful, i.e.: • At least one starter is successful • The main dish is successful • The dessert is successful In other words, if there's a successful starter, a successful main dish, and a successful dessert, then the group wins even if one starter has failed. But if the main dish has failed or the dessert has failed, then the group has failed the task regardless of the success of the other dishes. The four chefs Oren, Benni, Gidi and Doron participate in one of the shows. After receiving their task, they decided to split the preparation between them so that each chef prepares one of the four dishes. The chefs did not agree on who will prepare which dish, so they decided to determine it by chance. It turned out that Oren prepares a starter, Benni prepares a starter, Gidi prepares the main dish, and Doron prepares the dessert.
1. How much responsibility, do you think, does each of the chefs have for the success or failure of the task? 2. The show panel has tried the dishes and determined that none of the dishes was successful. Therefore the group has failed the task. To what extent, do you think, is each of the group members to blame for the group's failure? → **all fail** 3. To what extent, do you think, would each of the group members be to blame had it been determined that Gidi's main dish was successful, whereas the other three dishes were not? → ***B* succeeds** 4. To what extent, do you think, would each of the group members be to blame had it been determined that Benni's starter was successful, whereas the other three dishes were not? → ***C* succeeds**

4.2.2.1 Methods

Participants Eighty-three education undergraduate students from The Hebrew University of Jerusalem were recruited at the end of class and participated for course credit.

Materials and procedure All participants received identical forms that included the scenario depicted in Table 4.6. Each question was followed by four 7-point Likert scales. Each scale was labeled by a name ('Oren', 'Benni', 'Gidi', and 'Doron'), with the end points of the scales labeled as 'not at all' (1) and 'very much' (7). Question 1 was presented below the scenario, whereas Questions 2–4 were presented on the back of the page with their order counterbalanced between participants. Since no effect was found for the order of presentation the responses were aggregated across orders. Participants

were instructed to respond to Question 1 before turning the page and not to change their response after reading the subsequent questions.[17]

4.2.2.2 Results and Discussion

The blame attributions obtained for Questions 2–4 are presented in Figure 4.7. In order to test whether participants' blame attributions differed between the players and situations, we conducted a repeated-measures ANOVA with Player (A, B, C and D) and Situation (all failed, B succeeded, C succeeded) as within-subjects factors. We found main effects of Player, $F(3, 204) = 63.60, p < .001, {\eta_p}^2 = .483$ and of Situation, $F(2, 136) = 60.63, p < .001, {\eta_p}^2 = .471$ as well as an interaction effect, $F(6, 408) = 76.34, p < .001, {\eta_p}^2 = .529$. Having established that participants' blame attributions were influenced by our experimental manipulation, we proceed with a series of pairwise t-tests to test the more specific comparisons for which the models discussed above make different predictions.[18]

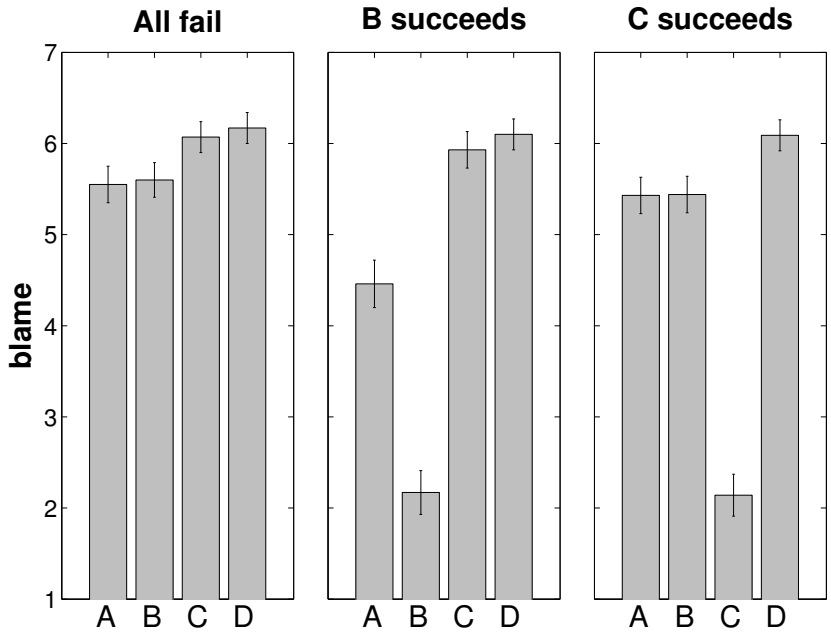


Figure 4.7: Mean blame attributions in Experiment 1 to the four players A, B, C and D for the situations in which *all fail*, *B succeeds* and *C succeeds*. *Note*: Error bars indicate ± 1 SE.

The blame attributed to A is affected by the individual outcomes of the other team

[17] Question 1 was included as part of a separate research program. It is not analysed in the current experiment.

[18] Somewhat surprisingly, the average blame attributed to players who succeeded in their individual tasks was higher than the minimal possible value of 1. We attribute this result to a number of unmotivated participants who provided random responses (e.g., marking the same value throughout the form). This behaviour is close to absent in the better-controlled Experiment 2, which replicates all the qualitative results of the current experiment.

members. Compared to the baseline when all team members fail, blame is decreased when B succeeds (4.46 vs. 5.55, $t(81) = 4.288$, $p < .001$), thereby rejecting SimResp and SimPiv. Furthermore, the blame attributions depend not only on the number of team members who share the blame, but also on the causal relationships between them. A's blame is higher when C succeeds compared to when B succeeds (5.43 vs. 4.46, $t(82) = 3.910$, $p < .001$), thereby rejecting DiffResp. The best prediction is provided by CFPiv, although A's blame does not increase when C succeeds compared to when all four fail, contrary to the prediction of the model (5.43 vs. 5.55, $t(81) = -.709$, $p = .480$).

Another prediction of CFPiv not supported by the data is that all team members should receive the same blame when all fail. The minimal number of changes required for pivotality is identical ($N = 2$) for each player. For example, in order to render A pivotal, a counterfactual situation needs to be considered in which the values of C and D were changed from 0 to 1. Similarly, in order to render C pivotal, the values of A (or B) and D would need to be changed. However, players C and D, whose respective successes are necessary for a team win, are assigned more blame than players A and B ($F(3, 79) = 4.981$, $p < .005$). Furthermore, when C succeeds, D still receives more blame than A and B (6.09 vs. 5.43, $t(80) = 3.141$, $p < .005$ and 5.44, $t(79) = 3.336$, $p < .005$, respectively).

4.2.3 Experiment 2

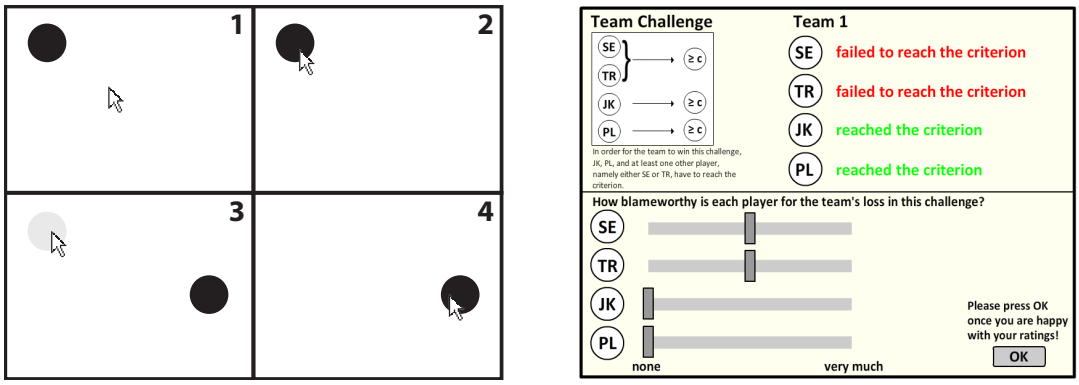
In contrast to the prediction of the CFPiv model, the blame assigned to A did not increase as the number of changes required to achieve the counterfactual situation in which A is pivotal decreased. To see whether blame does increase in the extreme case in which the required number of changes is reduced to zero, we added a new situation to Experiment 2, in which both C and D succeeded in their individual tasks (see Situation 4 in Figure 4.5b), thereby making A pivotal in the observed outcome. Additionally, we designed Experiment 2 to test the robustness of the results of Experiment 1 by repeating the same team challenge structure in a different framing, using computer interface, and with a different participant population.

4.2.3.1 Methods

Participants Sixty-one psychology undergraduate students at University College London participated in the experiment as part of a lab exercise.

Materials and procedure Participants were presented with the *dot-clicking game*, in which a dot is randomly repositioned on a computer screen each time the player clicks on it (see Figure 4.8a).[19] The score in the game is defined to be the number of clicks made

[19]Demos of Experiments 2 and 3 can be accessed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/finding_fault/finding_fault.html`



(a) The dot-clicking game **(b)** Screenshot of the Experiment

Figure 4.8: Diagram of the dot-clicking game and screenshot of the interface in Experiment 2.

within an fixed duration of time. In the experiment, hypothetical players in a team play the dot-clicking game. Each player succeeds in her game if she obtains a given minimal score. The team outcome is determined by a combination of the individual outcomes, which was presented graphically to the participants (see Figure 4.8b for a screenshot of the experiment). The structure of the game is equivalent to that used in Experiment 1. At the beginning of the experiment, participants played the dot-clicking game themselves to get a sense for the task. To avoid participants forming expectations based on their own performance, we stated that the game played by the hypothetical players was played for a different duration with a different-sized dot.

The first stage of the experiment was part of a separate study, and involved the participants making criticality attributions to players in different team challenges before the challenges are played. In the second stage of the experiment, participants used on-screen sliders to assign blame to the four players, in response to the following question: "How blameworthy is each player for the team's loss in this challenge?" The sliders corresponded to 11-point Likert scales (0 = 'not at all', 10 = 'very much'). The four different outcome patterns were presented in random order.

4.2.3.2 Results and discussion

The blame attributions are presented in Figure 4.9. As in Experiment 1, a repeated-measures ANOVA with Player and Situation as within-subject factors revealed significant main effects of Player, $F(3, 180) = 32.69, p < .001, {\eta_p}^2 = .353$ and of Situation, $F(3, 180) = 58.58, p < .001, {\eta_p}^2 = .494$ as well as an interaction effect, $F(9, 540) = 231.60, p < .001, {\eta_p}^2 = .794$. The patterns of blame attributions fully replicate those observed in Experiment 1. Compared to the baseline condition in which all players failed, the blame attributed to A significantly decreases when her substitute B succeeds (3.69 vs. 5.75, $t(60) = 4.972$, $p < .001$), but does not change significantly when

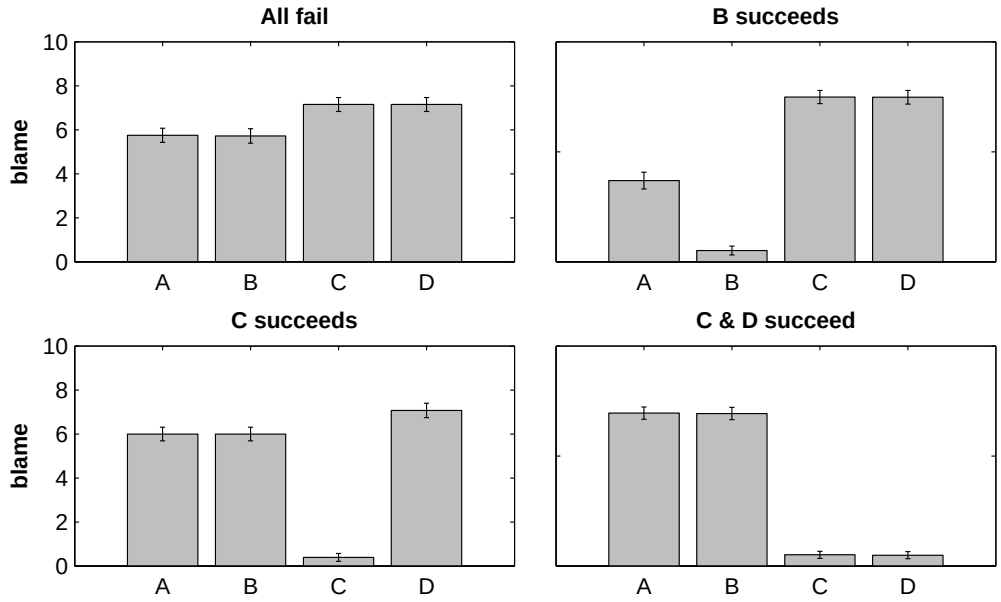


Figure 4.9: Mean blame attributions in Experiment 2 to the four players A, B, C and D for the situations in which *all fail*, *B succeeds*, *C succeeds* and *C and D succeed*. *Note*: Error bars indicate ± 1 SE.

her complement C succeeds (6.00 vs. 5.75, $t(60) = 1.023$, $p = .310$). However, when both complementary players C and D succeed, so that A becomes pivotal, A incurs significantly more blame (6.95 vs. 5.75, $t(60) = 3.687$, $p < .001$). As in Experiment 1, when all four team members have failed their individual tasks, then players C and D are perceived more blameworthy than A and B ($F(3, 180) = 18.435$, $p < .001$). Similarly, D is assigned more blame than both A and B when only C succeeds (7.07 vs. 6.00, $t(60) = 2.833$, $p = .006$ for either comparison).

Taken together, the results of the two experiments establish that blame attributions made by our participants are sensitive to the causal structure. The highest blame is assigned to an agent in the situation in which she was pivotal, and the lowest blame is assigned when the most changes of individual outcomes are required in order to make the agent counterfactually pivotal. In contrast with diffusion of responsibility considerations, reducing the number of agents who share the blame has different effects when different agents' outcomes are changed, and can even reduce the blame, depending on the causal structure and the relationship between the players. Thus, out of the four models we consider, the model of counterfactual pivotality provides the best explanation of the blame attributions observed in the experiments so far.

However, several findings remain unexplained by the model. In both experiments, the success of player C was not sufficient to increase the blame attributed to A, as predicted by the model, although the predicted effect was obtained in Experiment 2 when both C and D succeeded. None of the theoretical considerations can explain the lack of effect in the former case, as it is predicted by both the CFPiv and DiffResp

models. More interesting is the systematic difference in blame between players A and B on one hand and C and D on the other hand within the same situation, when the minimal number of changes required to make an agent pivotal is the same for all those who failed in their individual task. We conjecture that this result could be explained by the following observation: in the situation in which all of the team members have failed, the minimal change required to make each of them pivotal involves changing the outcomes of two other members. However, there is only one way to achieve this for team members A and B, namely by counterfactually changing the individual outcomes of C and D. In contrast, there are two ways to make each of C and D counterfactually pivotal, namely by changing the outcome of the other one as well as that of *either* A or B. For example, to make D counterfactually pivotal one must change C as well as either A or B. The same rationale holds for the situation in which only C succeeded. In this case, to make D counterfactually pivotal one must change either A or B, whereas to make A counterfactually pivotal one must change D.

This explanation implies that a minimal change model does not reflect the way in which people make responsibility attributions. Rather, when multiple paths exist in which an agent can be made counterfactually pivotal, blame increases accordingly. In the following section we develop and test a model that expands the CFPIv model to incorporate this insight.

4.2.4 Multiple counterfactual pivotality

Consider the following situation: You are the manager of your country's team in the International Salsa Competition (see Figure 4.10). Your team consists of Alice, Bob, Chuck and Dan. In order to compete in the tournament, Alice will need to show up and at least one of her partners. You instruct all of them to come to the tournament. However, as it turns out, none of them show up on the day of the competition. How much would you blame Alice for the fact that your team could not compete? How much would you blame Bob, Chuck or Dan?

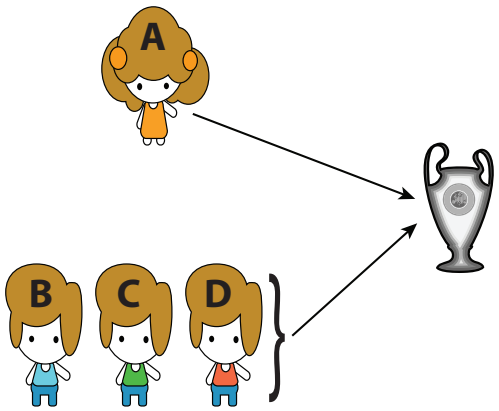


Figure 4.10: The dancing competition.

The CFPiv model predicts that all of them will be blamed equally. Given that none of them showed up, a minimum of one change needs to be made in order to render Alice pivotal. We can either change Bob, Chuck or Dan from *not having showed* up to *having showed up*. Similarly, for Bob, only one change is needed to render him pivotal, namely changing Alice to having showed up. The same holds of course for Chuck and Dan. Hence, we see that all team members are predicted to receive equal blame. However, the intuition is strong that Alice carries more blame for the fact that the team could not compete than each of her partners, as there are more counterfactual situations in which her appearance is crucial for the team to compete.

In this section, we introduce a new model, which we term *multiple counterfactual pivotality* (MultCFPiv). The new model expands CFPiv to account for the results of Experiments 1 and 2. Recall that the CFPiv model assigns responsibility according to the minimal change required to attain pivotality. Our new model retains the principle of counterfactual pivotality, but allows for multiple counterfactual situations, in which an agent is pivotal, to be considered. The new model has three important features. First, adding new paths by which an agent can become pivotal increases her responsibility. Second, as in the CFPiv model, responsibility decreases with the number of changes required to attain pivotality along any single path. Lastly, the new model reduces to the CFPiv model if there is only one way in which the agent can become counterfactually pivotal.

In order to accommodate multiple paths to pivotality while maintaining the general framework specified by the CFPiv model, we define an *equivalent single path* for any situation in which multiple paths exist. A path in this context is simply defined as a series of changes to the individual outcomes of other team members required to turn the observed situation into a counterfactual situation in which the target agent is pivotal. The responsibility assigned to the agent in the multiple-paths situation is the same as that assigned by the CFPiv model with the equivalent single path. The number of changes, N, is defined to be 0 if the agent is already pivotal and otherwise

$$N = \frac{1}{\sum_{i=1}^{k} \frac{1}{n_i}}, \tag{4.2}$$

where k is the number of different paths by which the agent can become pivotal, with required number of changes $n_1, n_2, \ldots, n_k$, respectively. The responsibility can then be defined to be $1/(N+1)$, as in the original CFPiv model.[20]

It remains to define how the number and lengths of the multiple paths are determined based on the causal structure of the team challenge and the individual players'

[20] This definition relies on the harmonic mean of the number of changes, and mirrors similar equivalencies in physical systems such as hydraulics and electricity. For example, it is isomorphic to the resistance in an electric circuit, in which each change is represented by a resistor of 1Ω, and resistors are connected serially to represent the number of changes along a path, and in parallel to represent multiple paths. We thank Yaniv Edery for suggesting this analogy.

outcomes. The first step is to identify all of the counterfactual outcome profiles in which an agent would be pivotal, and to determine the differences between each such counterfactual situation and the actually observed situation. Note that ordering the sequences sequentially defines a series of changes, or a path, that turns the actual situation into the counterfactual one. Next, exclude the situations for which the target agent is pivotal at an earlier step along one or more paths.

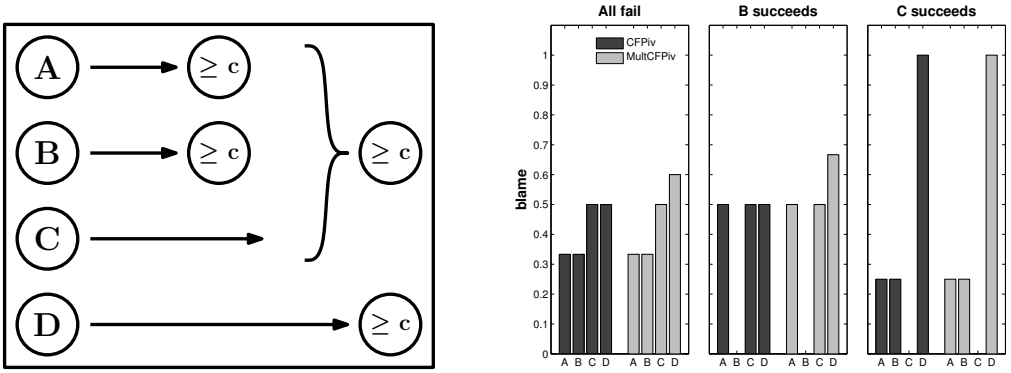
To illustrate, consider the team challenge in Figure 4.5a. As noted earlier in the discussion of our experimental results, if all four team members have failed their individual tasks, it is possible to make any of them pivotal through two counterfactual changes. The blame assigned to any team member according to CFPiv is $\frac{1}{2+1} = \frac{1}{3}$. For both A and B, there exists exactly one counterfactual situation in which they are pivotal, hence MultCFPiv also assigns them a blame of $\frac{1}{3}$. Conversely, for either C or D, there are three counterfactual situations in which they are pivotal. Namely, when the other one succeeds, in addition to either A, B, or both A and B. Since any path to the latter situation (in which both A and B succeeded) must go through one of the first two (in which either A or B succeeded), we exclude it from the analysis. Thus, we end up with two paths by which pivotality can be reached, each involving two changes. The number of changes in the equivalent single path is given by $\frac{1}{\frac{1}{2}+\frac{1}{2}} = 1$. Therefore, the responsibility assigned to C and D by our model is $\frac{1}{1+1} = \frac{1}{2}$.

As a further illustration, consider the team challenge in Figure 4.11a, in which the team wins if D succeeds in addition to either C or both A and B, hence $t = min(max(o_A, o_B), o_C) \times o_D$. Once more, assume that all four team members have failed in their individual tasks. In this case, the CFPiv model assigns a blame of $\frac{1}{3}$ to A and B and $\frac{1}{2}$ to C and D. For example, to make A pivotal we need to change B and D, so $N = 2$, to make D pivotal we need to just change C, so $N = 1$, and to make C pivotal we just need to change D, so $N = 1$. The predictions of MultCFPiv differ only with regard to team member D, who is pivotal in three counterfactual situations, namely when C succeeds, when A and B succeed or when A, B and C succeed. As in the previous example, the model does not consider the latter situation in which all three other team members have succeeded, since a subset of the changes is sufficient for pivotality. There remain two paths to pivotality. One is by changing the outcome of C ($n_1 = 1$), the other is by changing the outcomes of both A and B ($n_2 = 2$). The number of changes in the equivalent single path is now $\frac{1}{\frac{1}{1}+\frac{1}{2}} = \frac{2}{3}$, and the blame assigned to D is hence $\frac{1}{\frac{2}{3}+1} = 0.6$. Experiment 3 tests the novel prediction that team member D incurs more blame than the other three in the team challenge of Figure 4.11a, in the case that all four team members failed their individual tasks.

4.2.5 Experiment 3

To test the novel predictions derived from the MultCFPiv model, we constructed the team challenge depicted in Figure 4.11a, in which the team wins if D succeeds in addition

to either C or both A and B (i.e. $t = max(min(O_A, O_B), O_C) \times O_D$). The new challenge also serves as an additional test of the hypotheses tested in the previous experiments. We argued above that an implication of the CFPiv model is that how much blame a player incurs, reduces with each successful substitute and increases with each successful complement. The new team challenge provides a test for this generalization. In this challenge, player A is complementary to player B, and is a substitute of player C. As in the previous experiments, we start with a baseline situation in which all team members failed in their individual task, and compare blame attributions to player A when we reduce the number of failed team members. As in the previous team challenge (cf. Figure 4.5a), the failure of player D ensures that none of the other players is pivotal.



(a) Team structure. **(b)** Predicted blame attributions.

Figure 4.11: (a) Team structure and (b) predicted blame attributions according to the Counterfactual Pivotality (CFPiv) and the Multiple Counterfactual Pivotality (MultCFPiv) in Experiment 3.

Figure 4.11b presents the blame attributions predicted by the CFPiv and MultCFPiv models for each player in each of the experimental conditions. A comparison of the bars in the figure reveals the qualitative predictions tested in the experiments. The basic prediction of both models is tested by comparing the blame assigned to player A in the three conditions, as in the previous experiments. Namely, A receives more blame if B succeeds, but less blame if C succeeds. The two models differ with regard to the blame attributed to player D. The prediction of the MultCFPiv model to be tested is that D is more to blame than C when all fail or B succeeds, although both can become pivotal through only one change. In addition, the success of B reduces the number of changes required to make D pivotal along the alternative path, which is ignored in CFPiv, hence only MultCFPiv predicts a higher blame for D as a result.

4.2.5.1 Methods

Participants Forty participants from the USA, 13 males and 27 females, ages 19–57 ($M = 32$) were recruited to participate in the experiment via Amazon Mechanical Turk for a flat fee of $1 (see Mason & Suri, 2012, for a discussion of conducting behavioural

research on Amazon's Mechanical Turk).

Materials and procedure The procedure was similar to that of Experiment 2. The team challenge used was the one depicted in Figure 4.11a, and the three conditions were (a) all fail, B succeeds, and C succeeds. Participants provided blame attributions on sliders corresponding to 21-point Likert scales.

4.2.5.2 Results and discussion

Generally, the patterns observed in the previous two experiments were replicated with the new team challenge (see Figure 4.12). A repeated-measures ANOVA with Player and Situation as within-subject factors revealed significant main effects of Player, $F(3, 234) = 89.52, p < .001, {\eta_p}^2 = .534$ and of Situation, $F(2, 156) = 50.96, p < .001, {\eta_p}^2 = .395$ as well as an interaction effect, $F(6, 468) = 62.64, p < .001, {\eta_p}^2 = .445$. The blame assigned to player A significantly decreases if the substitute player C succeeds (8.13 vs. 13.00, $t(39) = 4.452$, $p < .001$), but does not significantly differ if the complement player B succeeds (12.08 vs. 13.00, $t(39) = -1.045$, $p = .303$). The effect of reducing the number of team members who failed, significantly depends on the role of the team member who succeeded in the individual task, as player A receives more blame when B succeeds compared to when C succeeds (12.08 vs. 8.13, $t(39) = 3.209$, $p < .005$).

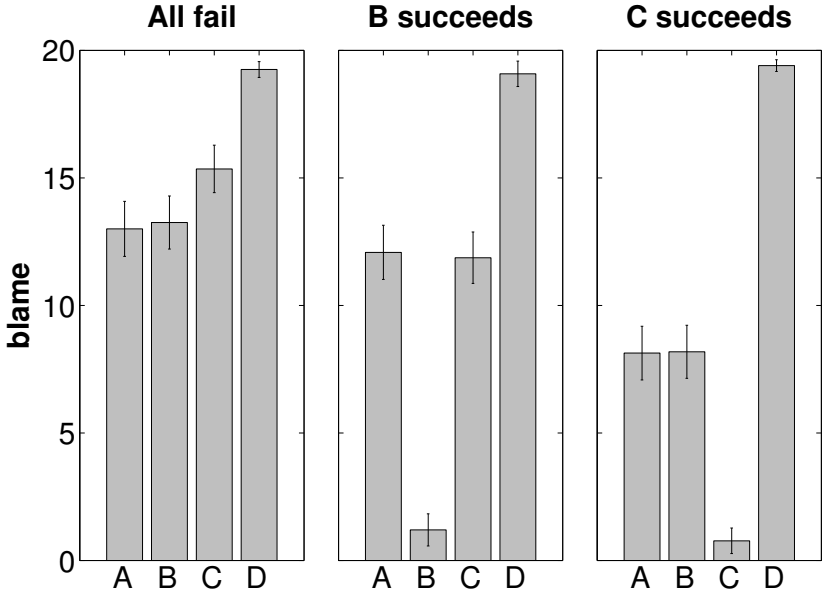


Figure 4.12: Mean blame attributions in Experiment 3 to the four players A, B, C and D for the situations in which *all fail*, *B succeeds* and *C succeeds*. *Note*: Error bars indicate ± 1 SE.

The new challenge produces new test cases for the MultCFPiv model. In the situations where the CFPiv and MultCFPiv models diverge, the results are in line with

MultCFPiv. When all of the four players fail, player D receives more blame than player C (19.25 vs. 15.35, $t(39) = 4.444$, $p < .001$), who in turn receives more blame than player A (15.35 vs. 13.00, $t(39) = 3.116$, $p < .005$) or player B (15.35 vs. 13.35, $t(39) = 2.723$, $p < .01$). Similarly, player D is perceived as more blameworthy than player C when player B succeeds (19.08 vs. 11.87, $t(39) = 7.232$, $p < .001$). However, the blame attributed to player D does not significantly change across situations, possibly due to a ceiling effect, as the average blame rating is above 19 out of 20 in all three situations.

The data yield one surprising result which is not predicted by any of the models we consider. Player C is assigned less blame when player B succeeds compared to when all of the players fail (11.87 vs. 15.35, $t(39) = 3.575$, $p < .001$). Note that the relationship between players B and C is one of substitution. Hence the finding, albeit not predicted by MultCFPiv, is consistent with the general principle implied by counterfactual pivotality reasoning which states that responsibility is reduced when a peer succeeds in the case of substitution.

In sum, the results of the new experiment are consistent with the results obtained with the previous team challenge. Out of the five models we consider, the model based on multiple counterfactual pivotality explains the data best. Although some differences predicted by the model are not apparent in the data, we take the results to confirm the basic role of counterfactual pivotality in blame attributions.

4.2.6 General discussion

In this section we have seen a simple and clear test for possible models designed to capture the way in which people make responsibility attributions in a team environment. These relationships are apparent in all of our experiments. The effect of a change in one team member's performance on the blame incurred by her peer strongly depends on the way in which the respective contributions of the two interact with regard to the team outcome.

Our results enabled us to extend the CFPiv model tested by Gerstenberg and Lagnado (2010). The CFPiv model only takes into account the minimal number of changes along a single path to render the person under consideration pivotal. In contrast, the MultCFPiv model is sensitive to how many paths there are to reach a counterfactual situation in which the person would be pivotal and how many changes to the actual situation would be required along each path.

The sensitivity to counterfactual causal reasoning implied by the experimental results can be interpreted in different ways. The number of paths and counterfactual changes at the heart of the model can be taken as mental steps or, alternatively, as reflecting the difficulty of bringing to mind a certain counterfactual state given the actual state. Thus, the model need not be taken as a literal process model, but as a support for the

importance of counterfactual causal reasoning in responsibility attributions in group contexts.

4.3 Pivotality and Criticality (Lagnado et al., accepted)

The results from Gerstenberg and Lagnado (2010) and Zultan et al. (2012) confirm that the structural approach to responsibility successfully explains attributions to players who, in retrospect, would not have altered the team outcome. It achieves this by using a graded notion of responsibility determined by the number of changes that are needed to render a player pivotal for the outcome. The approach does not allow, however, for graded attributions of responsibility with regard to players who are already pivotal. Nonetheless, certain cases generate a strong intuition that pivotal individuals in different situations are not regarded as equally responsible. Consider the classic bystander effect (Darley & Latané, 1968): a victim is attacked by an offender and in need of help. Several observers are in a position to potentially intervene. As is well-known, people have a reduced sense of responsibility when there are others who could also help compared to a situation in which there is no one else. However, note that the structural model does not predict this effect. The structure of the situation is disjunctive (assuming that one person would be enough to fend off the offender) and thus each person is pivotal for the negative outcome, irrespective of the number of people present.

This intuition can be accommodated within the structural framework by incorporating the concept of *prospective responsibility* (see Cane, 2002; Hamilton, 1978; Schlenker et al., 1994): the extent to which a person is perceived to be critical for an outcome *before* it has occurred. Accordingly, a reason for why a single person is held more responsible for not helping a stranger than each individual in a group of people, is that the single person could have prevented the outcome (he was pivotal) *and* he knew that the outcome only depended on him (he was more critical). *Retrospective* responsibility might not only be influenced by whether or not a person was pivotal *after* the fact but also by how critical a person was perceived to be *prior* to the outcome.

As briefly outlined in Chapter 2, it has already been argued that responsibility attributions are not only sensitive to what a person actually did but also to what role the person had. Hamilton (1978), for example, claimed that attributions of responsibility are markedly influenced by what a person was supposed to do. In a similar vein, Schlenker et al. (1994) argue that responsibility "acts as a psychological adhesive that connects an actor to an event and to relevant prescriptions that should govern conduct" (Schlenker et al., 1994, p. 632).

Their *Triangle Model of Responsibility* features three core constructs: (1) the *prescriptions* relevant in a particular situation (similar to the social norms we have discussed above), (2) the nature of the *event* that has taken place and (3) the *identity* of the agent. Their model predicts that responsibility increases with a perceived increase

of the strength in the linkages between the three core constructs. That is, an agent will be judged more responsible if salient and non-ambiguous prescriptions of conduct were applicable to the event (prescription-event link), if the prescriptions clearly apply to the agent's role (prescription-identity link, cf. Hamilton, 1978) and if the actor is connected to the event in an appropriate way, for example, through having sufficient control over the outcome of the event (identity-event link). Interestingly, Schlenker et al. (1994) argue that not only will responsibility increase as a function of the strength of the linkages but also what the authors refer to as 'determination'. With determination, they mean the actor's perseverance in a task and the commitment to achieving her goals. "We propose that *before a performance*, the strengths of the three linkages, in combination with the potency of the elements, will affect the actor's determination to achieve prescribed goals." (Schlenker et al., 1994, p. 637, emphasis added).

As we have seen in Chapter 2, Weiner (1995) has also claimed that there is an intimate relationship between causality, responsibility and motivation. Indeed, Kerr and Bruun (1983) have found that different causal structures influence how much effort individuals in a group are willing to exert and how critical they perceive their contribution to be for the group's success. They experimentally varied the abilities of the group members and the nature of the task structure (conjunctive vs. disjunctive). They found that effects of member ability and group structure interacted. In disjunctive tasks, participants with *low* ability reduced their efforts whereas in conjunctive tasks participants with *high* ability tried less hard. To get more direct evidence for their hypothesis that motivation is linked to perceptions of criticality, participants were asked after the experiment to rate how much they thought the group success depended on their own performance. In these ratings, a qualitatively identical pattern to the actual exertion of effort was observed. More able participants felt more critical in disjunctive tasks compared to conjunctive tasks. Conversely, weaker participants rated themselves as more critical in conjunctive as opposed to disjunctive tasks (see also Hertel, Kerr, & Messé, 2000; Kerr, 1983).

Relatedly, there is direct empirical evidence that the degree to which an individual perceives their action to be critical for the group outcome affects whether they are willing to cooperate in a public goods game (Au, 2004; Kerr, 1989, 1992; Kerr & Kaufman-Gilliland, 1997; Rapoport, Bornstein, & Erev, 1989). Some researchers have argued, that cooperative behaviour is directly influenced by the degree to which individuals perceive themselves responsible for the others in the group (Berkowitz & Daniels, 1963; De Cremer & van Dijk, 2002; De Cremer & Van Lange, 2001; Fleishman, 1988; Kerr, 1995).

Legal scholars (e.g. Cane, 2002; Hart, 2008) have also argued for a close relationship between *prospective* and *retrospective* responsibility. Retrospective responsibility refers to the degree to which individuals are to be held accountable for negative consequences that have resulted from their actions. These considerations form the basis

for punishment. Prospective responsibility is related to positive and negative duties that a person has as part of their role. It describes the responsibility a person has for preventing negative outcomes (e.g. a security guard) and producing positive ones (e.g. a politician).

In this section, we show that incorporating the notion of criticality explains additional deviations from the predictions of the structural model (Chockler & Halpern, 2004) and present a new experiment to test the respective roles of criticality and pivotality in intuitive judgments of responsibility. The new framework subsumes the 'multiple paths to pivotality' model introduced and tested in Zultan et al. (2012).

4.3.1 Models of criticality

Let us summarise the key intuitions with respect to criticality: an individual's perceived criticality diminishes with an increased number of people in disjunctive structures because the actions of each individual in the group are alone sufficient to bring about the outcome. However, an individual's criticality is not expected to decrease with an increased group size for conjunctive tasks. Regardless of the number of conjuncts, each individual's action is still necessary for the team outcome. To illustrate, imagine that a victim is attacked by three offenders and three observers are present each of whom would need to intervene in order to help the victim. Here, the intuition is that each of the observers is highly critical for preventing the outcome (and more critical than in a situation in which only one offender was present but there were still three observers).

So far we have relied on intuitive perceptions of criticality and have not given a formal model of criticality. We now consider different ways in which these intuitions can be formalised and generalised. This will allow us to define a criticality-pivotality framework to be tested in a new experiment. We consider two possible models of criticality, which we shall refer to as the *expected pivotality model* and the *heuristic model*. These models are intended as an example for how criticality can be formalised and incorporated into a model of retrospective responsibility. Other approaches to criticality may be just as valid as the ones we discuss here. We will briefly mention one such alternative approach when discussing the heuristic model.

4.3.1.1 Expected pivotality model

Rapoport (1987) provided a definition of the criticality of a person's contribution in the step-level public goods game. In this game, each person in a group is endowed with an initial amount of money, let's say \$5. Each person then indicates whether they want to contribute their money to the public good. If a sufficient number of people provide their endowment and the provision point is met, the public good is provided (and, for example, each person gets an additional \$10). The public good is available to all in the group, no matter whether or not they contributed their endowment. According to

Rapoport's (1987) definition, a player's criticality is given by the probability that their contribution will make a difference to the outcome. In the step-level public goods game, this is the probability that the number of contributors among the other players is exactly one less than the number required to provide the public good, given prior beliefs about contributions in the group. Formally, the criticality of player A in situation S is given by

$$\pi_{pivotal}(A, S) \;=\; \binom{n-1}{k-1}\, p^{k-1}\, (1-p)^{n-k}, \tag{4.3}$$

where n equals the number of players in the group, k denotes the provision point that has to be reached in order for the public good to be provided and p is given by the probability that another player in the group will contribute.[21] For example, in a public goods game with $n = 5$ players, a provision point of $k = 3$ and a probability of $p = 0.6$ that each of the other players will contribute, the probability that the player under consideration will be pivotal is $\pi_{pivotal} = \binom{4}{2}\, 0.6^2\, 0.4^2 = 0.35$.

Does this definition of criticality as 'expected pivotality' capture our intuitions with regard to the examples described above? It correctly predicts that an individual's criticality reduces with an increased group size in disjunctive situations. For example, whereas a single person's action will always be pivotal, the probability that a person's action will be pivotal decreases with an increased group size. In the *single offender case*, an individual's decision to help will only be pivotal if none of the others acts. However, contrary to intuition, the model also predicts that criticality will reduce *in the same way* for conjunctive tasks. Thus, holding the group size of three constant, the model predicts that each individual's criticality is identical in *three offenders case* and in the *single offender case*. Whereas an individual is pivotal if none of the others act in the *single offender case*, he is pivotal in the *three offenders case* only if both of the others act. Assuming that a person is maximally uncertain about whether the others will act or not, his expected pivotality is the same in both situations.

4.3.1.2 Criticality heuristic

The second model of criticality we consider is a simple *heuristic model*. This model assigns full criticality to players whose success is necessary for the team outcome and divides the criticality equally between players who share a task in a disjunctive fashion. This is best illustrated using the asymmetric task structure used above and reproduced in Figure 4.13.

In this challenge, both C and D have to succeed in their individual tasks in addition to at least one player out of A and B. The heuristic model assigns full criticality to C

[21] Note that this definition of criticality rests on the *homogeneity assumption*, which states that all other players will contribute to the public good with the same probability. See Rapoport (1987) for a definition of criticality that relaxes this assumption and allows different players to have unequal probabilities of contribution.

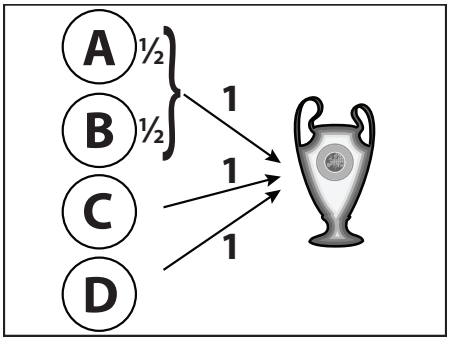


Figure 4.13: Asymmetric team challenge with predictions of the *heuristic criticality model*.

and D, whereas A and B only receive a criticality of $\pi_{heuristic}(A, S) = .5$ each, because their contributions combine in a disjunctive manner.

Another way to formalise the importance of a player whose success is necessary for the team to win is to model criticality as the relative decrease in the probability of the team winning when the player fails. This notion of criticality can be defined as $1 - \frac{p(win|fail)}{p(win|success)}$. According to this definition, a player's criticality varies from zero (when he has no effect on the team outcome) to one (when he is necessary for the team to win). The predictions of this model for the deterministic situations to be explored in the following experiment are virtually identical to the predictions of the heuristic model. We will, therefore, not discuss it separately in the following, but present it here as an illustration for how models of criticality can be extended to encompass non-deterministic situations.

4.3.2 Models of pivotality

We will consider two models of pivotality: the *counterfactual model* ($\rho_{counterfactual}$) which assigns 1 to pivotal players and 0 to non-pivotal players and the *structural model* ($\rho_{structural}$) which assigns degrees of responsibility as a function of the minimal number of changes that are required to make the player pivotal.

As a reminder, Figure 4.14 shows the predictions of the *structural model* for four simple situations. As outlined above, the model predicts that in a situation in which two players failed their tasks, A's responsibility is 1 in a disjunctive challenge and 1/2 in a conjunctive challenge. Conversely, the model predicts that when both players succeeded, A's responsibility is 1/2 for disjunction and 1 for conjunction.

4.3.3 Testing the criticality-pivotality model

We aim to test a general framework which predicts that participants' responsibility attributions are not only affected by how close a player's contribution was to being pivotal but also by how critical this player's contribution was perceived to be for the group outcome. Formally,

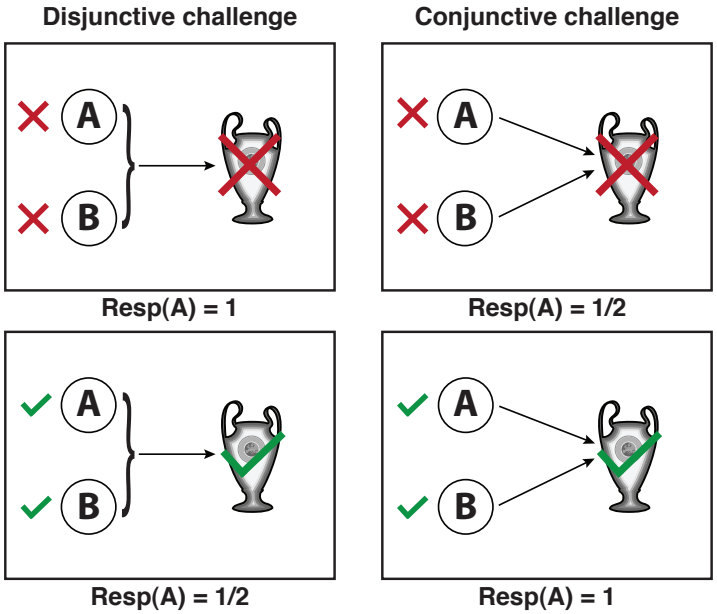


Figure 4.14: Predictions of player A's responsibility by the Structural model (Chockler & Halpern, 2004) as a function of outcome (failure vs. success) and task structure (disjunctive vs. conjunctive). *Note*: The curly braces indicate that players' performances combine in a disjunctive fashion; ✗ = failure, ✓ = success.

$$responsibility(A, O, S) = f(criticality(A, S), pivotality(A, O, S)) \tag{4.4}$$

where $criticality(A, S)$ denotes the criticality of player A in situation S and $pivotality(A,O,S)$ denotes A's pivotality for the outcome O in situation S. In the following, we illustrate how the general framework can be applied to make quantitative predictions about participants' responsibility attributions by using the different models of criticality and pivotality discussed above.

4.3.4 Experiment

We designed a new experiment to investigate the influence of both criticality and pivotality on participants' responsibility attributions. In the experiment, participants evaluated the performance of individuals in team challenges which differed in terms of group size and structure (see Figure 4.15). In the first phase of the experiment, participants judged how critical they perceived player A to be for the team outcome before the outcome was known. In the second phase, they judged how responsible player A was for the team outcome in different situations.

4.3.4.1 Methods

Participants 40 participants (25 female) aged 18–60 ($M = 33.86$, $SD = 11.76$) were recruited online via Amazon Mechanical Turk.

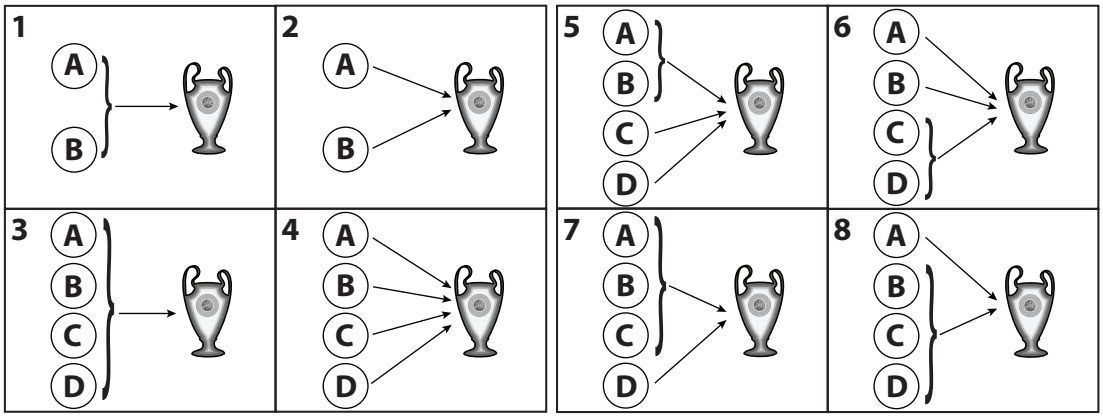


Figure 4.15: Team challenges used to investigate the effects of criticality and pivotality on responsibility attributions to player A.

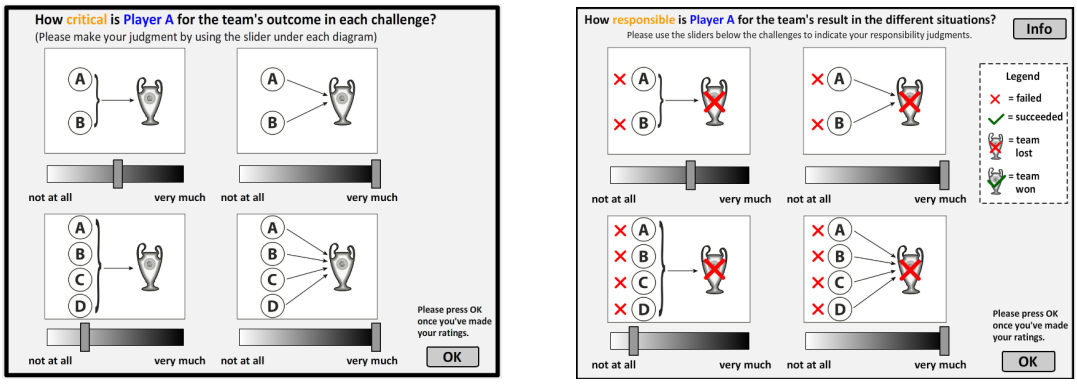
Materials and Procedure The experiment was programmed in Adobe Flash CS5.[22] Participants ran the experiment on their individual computers.

The experiment closely followed the procedure in Zultan et al. (2012). Participants were instructed that they will take the role of an external observer and that their task will be to evaluate the performance of contestants in a game show playing the dot-clicking game (see Figure 4.8a). Participants first played the dot-clicking game themselves. In order to prevent possible influences of participants' perceived difficulty of the task on their later attributions, participants were instructed that the contestants in the game show played a slightly different version in which the dot was smaller but they were given more time. After having played the game, participants were told that the contestants in the game show are randomly assigned to four different challenges (Challenges 1–4 in Figure 4.15). Participants learned via examples how the different team challenges worked. They were told that each challenge consisted of a new set of players.

Participants were then asked to judge how critical player A was for the team outcome in the challenges 1–4 (see Figure 4.15). Participants saw all four challenges on the same screen and answered the following question: "How critical is Player A for the team's outcome in each challenge?" (see Figure 4.16a) They made their judgments on separate sliders, which were positioned under the four different challenges. The endpoints of the sliders were labeled 'not at all' and 'very much'.

Participants then saw three sets of challenges for which they were asked to judge how responsible player A was for the team outcome. Participants saw the results of four different team challenges simultaneously on the screen (see Figure 4.16b) and answered the following question: "How responsible is Player A for the team's result in the different situations?" Again, they indicated their responses on separate sliders whose endpoints were labeled 'not at all' and 'very much'.

[22] A demo of the experiment can be accessed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/structure_demo.html`



(a) Screenshot of the *criticality* judgment phase

(b) Screenshot of the *responsibility* judgment phase

Figure 4.16: Screenshots of the experiment. (a) *Criticality* judgments *before* the outcome is known, (b) *Responsibility* judgments *after* the outcome is known. *Note*: ✗= failure, ✓= success.

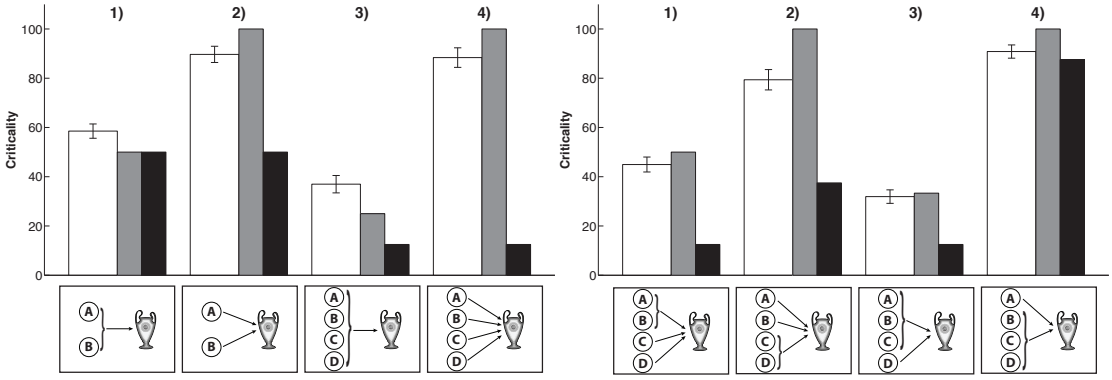
Afterwards participants were introduced to a set of asymmetric team challenges (see challenges 5–8 in Figure 4.15). Several examples ensured complete understanding of the novel structures. Participants then judged how critical player A was in the asymmetric structures before judging A's responsibility for the group outcome in six more sets of challenges. On average, it took participants 14.04 minutes ($SD = 4.99$) to complete the experiment.

Design The experiment contained two sets of four challenges for which participants rated A's *criticality* and nine sets of four challenges for which participants rated A's *responsibility*. The stimuli for the responsibility ratings were created to ascertain to what extent attributions of responsibility were influenced by criticality and pivotality. In some situations we varied pivotality but kept criticality constant. In other situations we varied criticality but held pivotality constant. Finally, some situations varied both criticality and pivotality.

4.3.4.2 Results

We will discuss participants' criticality and responsibility judgments in turn.

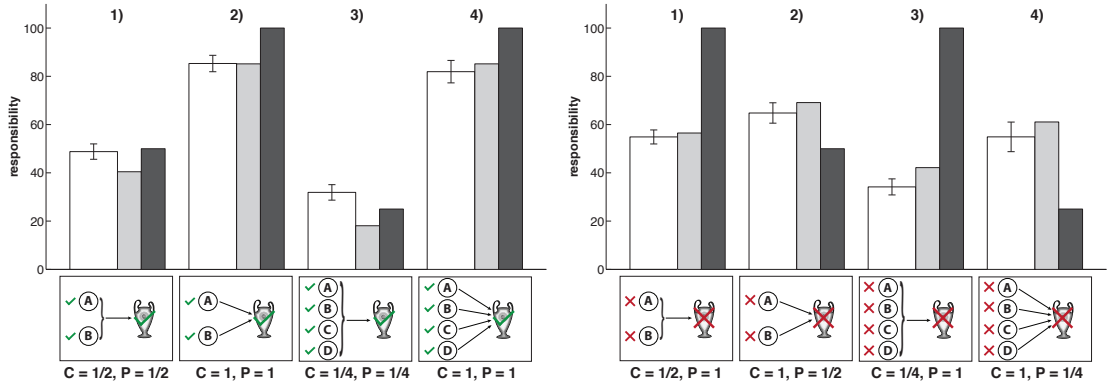
Criticality judgments Participants' criticality judgments are shown in Figure 4.17. The $\pi_{heuristic}$ model ($r = .97$, $RMSE = 11.15$) predicted participants' criticality judgments very well and better than the $\pi_{pivotal}$ model ($r = .62$, $RMSE = 37.42$). Generally, player A was rated highly critical when his individual success was necessary for the team win (see challenges 2 and 4 in Figure 4.17a and Figure 4.17b). When A formed part of a disjunctive (sub-)group, criticality reduced with the number of people in the group (see challenges 1 and 3 in Figure 4.17a and Figure 4.17b).



(a) First set of challenges **(b)** Second set of challenges

Figure 4.17: Mean criticality judgments for player A (white) and model predictions by the $\pi_{heuristic}$ model (grey) and the $\pi_{pivotal}$ model (black) for two different sets of challenges. Error bars indicate ± 1 *SEM*.

Responsibility judgments Figure 4.18 shows participants' responsibility judgments for two sets which contain only symmetrical challenges. The $\rho_{structural}$ model predicts participants' responsibility attributions well for the set of challenges shown in Figure 4.18a but not for the set shown in Figure 4.18b. Note, however, that in Figure 4.18a, criticality (as predicted by the $\pi_{heuristic}$ model which explains participants' criticality attributions best) and pivotality (as predicted by the $\rho_{structural}$ model) are perfectly confounded, whereas in Figure 4.18b, criticality and pivotality are negatively correlated. In this set of challenges, participants attributed more responsibility when A failed in conjunctive tasks versus disjunctive tasks in contrast to the prediction of the structural model. For example, although $N = 3$ changes would be necessary to make A pivotal when four players failed in the conjunctive challenge (Situation 4), A receives more re-

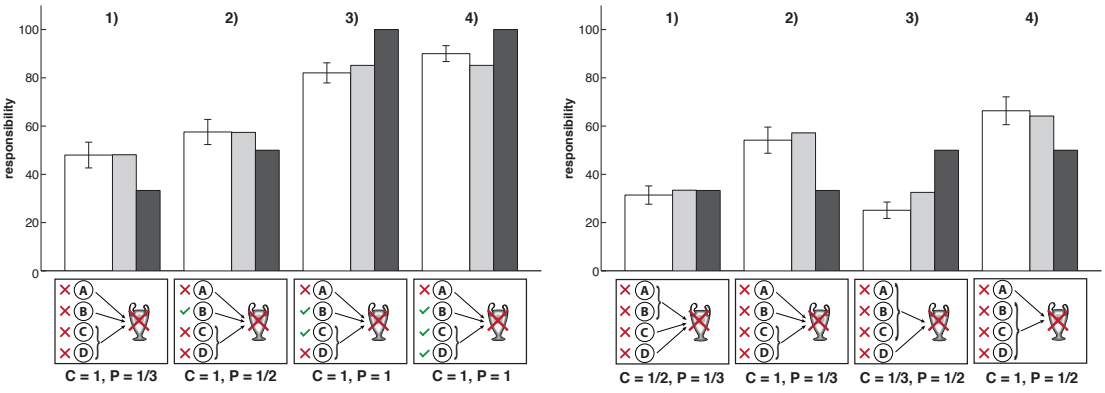


(a) All players succeeded **(b)** All players failed

Figure 4.18: Mean responsibility ratings (white bars) for two sets of challenges with the predictions of the *critical-pivotality model* (grey bars) and the $\rho_{structural}$ model (black bars). Note: C = Criticality, P = Pivotality. Error bars indicate ± 1 *SEM*.

sponsibility than in the situation in which four players failed in the disjunctive challenge (Situation 3), despite the fact that A is pivotal in this situation.

From this pattern of results, one could be led to infer that pivotality considerations are not essential for attributions of responsibility as participants' attributions closely follow the predictions of the criticality heuristic. However, our design also included situations in which A's criticality was held constant but his pivotality varied. If pivotality was not important, then A's responsibility should be the same in these situations. Consider the set of challenges shown in Figure 4.19a. Because the structure of the challenge did not vary between situations, A's criticality is equal in all four situations. However, participants attributed more responsibility to A when he was pivotal (Situations 3 and 4) compared to when he was not pivotal (Situations 1 and 2), $t(39) = 6.63, p < .001, r = 0.53$. In fact, the pattern of attributions closely followed the predictions of the $\rho_{structural}$ model: A's responsibility was lowest in Situation 1 ($M = 48$, $SD = 33.75$) in which $N = 2$ changes are required to make him pivotal. A's responsibility increased slightly in Situation 2 ($M = 57.55$, $SD = 32.92$) as N is reduced to 1. Responsibility further increased significantly in Situation 3 ($M = 82.05$, $SD = 26.34$) in which A was pivotal and only minimally in Situation 4 ($M = 90.03$, $SD = 20.82$).



(a) Criticality constant, pivotality varied

(b) Criticality varied, pivotality constant

Figure 4.19: Mean responsibility ratings (white bars) for two more sets of challenges with the predictions of the *critical-pivotality model* (gray bars) and the $\rho_{structural}$ model (black bars). Note: C = Criticality, P = Pivotality. Error bars indicate ±1 *SEM*.

The set of challenges shown in Figure 4.19b includes two pairs of situations in which A's pivotality was the same but her criticality different. $N = 2$ are required to make A pivotal in Situations 1 and 2. However, A is more critical in Situation 2 than in Situation 1. Participants held A significantly more responsible for the team's loss in Situation 2 ($M = 54.18$, $SD = 34.29$) compared to Situation 1 ($M = 31.43$, $SD = 23.98$), $t(39) = 6.01, p < .001, r = 0.48$. A's pivotality in Situations 3 and 4 is 1/2, however, A is more critical in Situation 4 than in Situation 3. A was seen as more

responsible in Situation 4 ($M = 66.35$, $SD = 36.41$) compared to Situation 3 ($M = 25.13$, $SD = 21.38$), $t(39) = 9.09, p < .001, r = 0.68$.

Together, the four sets of challenges discussed so far establish that responsibility attributions to individuals within a group are affected both by how critical the person was perceived to be and by how close she was to being pivotal. When criticality was held constant (see Figure 4.19a), responsibility increased with pivotality. When pivotality was held constant (see Figure 4.19b), responsibility increased with criticality. Thus, neither criticality nor pivotality alone are sufficient to explain responsibility attributions. This conclusion holds regardless of the particular models of criticality and pivotality used.

4.3.4.3 Modelling responsibility attributions

We first test how well simple models that don't combine criticality and pivotality can explain participants' responsibility attributions. As criticality models, we consider the *expected pivotality model* ($\pi_{pivotal}$) and the *heuristic model* ($\pi_{heuristic}$) described above. For models of pivotality, we consider a *simple counterfactual model* ($\rho_{counterfactual}$), which only assigns responsibility to pivotal players and the *structural model* ($\rho_{structural}$), which assigns responsibility as a function of the distance to pivotality.

The diagonal in Table 4.7 shows how well these simple models predict participants' responsibility attributions in the experiment. In general, the two pivotality models predict participants' attributions better than the two criticality models. More precisely, the $\rho_{structural}$ model predicts participants' attributions best followed by the $\rho_{counterfactual}$ model, the $\pi_{heuristic}$ model and the $\pi_{pivotal}$ model.

We also considered models which predicted responsibility as a weighted function of criticality and pivotality. Thus,

$$Responsibility(A) = \alpha + w \times criticality(A) + (1 - w) \times pivotality(A), \tag{4.5}$$

where α is a global intercept and w is a weighting parameter whose range is constrained from 0 to 1.[23]

The cells *off* the main diagonal in Table 4.7 show the model fits of these weighted models. Overall, a model that uses a combination of the $\pi_{heuristic}$ model for criticality and the $\rho_{structural}$ model of pivotality explains participants' attributions best. However, a model that replaces the $\rho_{structural}$ model with a $\rho_{counterfactual}$ model also explains the data well. This is not surprising, given that the predictions of both models are strongly correlated for the set of situations we employed ($r = .97$). As a check, we also included weighted models that combined two models from the same family (i.e. two *criticality*

[23]For the model predictions of the *critical-pivotality model* in Figures 4.18, 4.19, and 4.21 we used the $\pi_{heuristic}$ model for criticality and the $\rho_{structural}$ model as a model of pivotality. Furthermore, we allowed w to vary between the different sets of challenges.

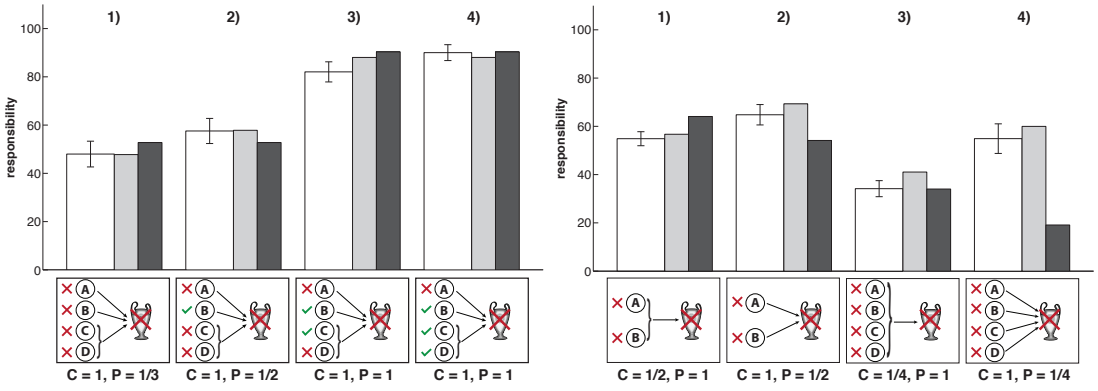
Table 4.7: Correlations of *criticality models* and *pivotality models* with participants' responsibility attributions. *Note*: Cells on the diagonal show predictions of simple models, cells off the diagonal show predictions of combined models.

	Criticality		**Pivotality**	
	$\pi_{pivotal}$	$\pi_{heuristic}$	$\rho_{counterfactual}$	$\rho_{structural}$
$\pi_{pivotal}$	.59 (24.44)	-	-	-
$\pi_{heuristic}$	.71 (19.57)	.67 (22.41)	-	-
$\rho_{counterfactual}$	.83 (16.12)	.89 (13.92)	.74 (36.14)	-
$\rho_{structural}$	.84 (13.66)	.90 (10.81)	.74 (16.55)	.77 (19.59)

Pearson correlations r with *RMSE* in parentheses.

models or two *pivotality models*). These weighted models perform much worse than the ones which combine one model from the criticality family with one model from the pivotality family.

Although the overall correlations of some of the combinations of models are similarly high, the specific sets of challenges we have discussed so far support a combination of the $\pi_{heuristic}$ model of criticality with the $\rho_{structural}$ model of pivotality. For example, the pattern of attributions in Figure 4.20a in which criticality was held constant, is only in line with the $\rho_{structural}$ model but not with the $\rho_{counterfactual}$ model. The $\rho_{counterfactual}$ model cannot predict that A's responsibility increases in situations in which he was not pivotal (Situation 1 vs. Situation 2). The same holds for the set of challenges



(a) $\pi_{heuristic} \times \rho_{structural}$ vs. $\pi_{heuristic} \times \rho_{counter}$

(b) $\pi_{heuristic} \times \rho_{structural}$ vs. $\pi_{pivotal} \times \rho_{structural}$

Figure 4.20: Comparison of different criticality-pivotality models. The best-fitting model, $\pi_{heuristic} \times \rho_{structural}$ (grey bars) is compared with (a) the $\pi_{heuristic} \times \rho_{counterfactual}$ model and (b) the $\pi_{pivotal} \times \rho_{structural}$ model (black bars). Error bars indicate ± 1 *SEM*.

shown in Figure 4.21b which implements the situations used in Zultan et al.'s (2012) Experiments 1 and 2. Similarly, participants' judgments shown Figure 4.20b cannot be explained if the $\pi_{pivotal}$ model is used as a model of criticality but it follows naturally from the $\pi_{heuristic}$ model. The $\pi_{pivotal}$ model cannot predict this pattern since it assigns equal criticality to players in disjunctive and conjunctive structures as discussed above.

Figure 4.21 shows the five remaining sets of challenges. We will not discuss these patterns of results in any detail, however, we would like to stress that each of the patterns of attributions within a given set of challenges is consistent with the *critical-pivotality model* which combines the $\pi_{heuristic}$ with the $\rho_{structural}$ model. In contrast, for each of the other possible combinations, there are qualitative patterns which cannot be explained.

Figure 4.22 shows a scatter plot of the model predictions and mean attributions for the 36 patterns employed in the experiment for the three best-fitting combinations of criticality and pivotality models. As described above, the $\pi_{heuristic} \times \rho_{structural}$ model fits participants' attributions best. However, it is worth noting that there is remaining variance in people's attributions that is not accounted for by the model. Whereas the model predicts that all cases in which a player was highly critical *and* pivotal are treated the same, participants differentiate between these situations (see rightmost part of the scatter plot in Figure 4.22a). We will discuss a potential explanation for this finding below.

4.3.5 Discussion

The overall results of the experiment show that participants' responsibility attributions are very well explained by assuming that they take into account both (i) how critical a person's contribution was for the group outcome, and (ii) whether or not a person's contribution would have made a difference to the outcome (and if not, how close it was to making a difference). We have seen that one concrete implementation of this general framework which uses the heuristic model as a model of criticality and the structural model as a model of pivotality explains participants' attributions particularly well. While there might be implementations that fit participants' attributions better, we are confident that any model that does a reasonably good job of predicting people's attributions will need to be sensitive to both criticality and pivotality.

In future research, we will aim to combine the ideas of criticality and pivotality into an integrated model of responsibility attribution. What we see as an important contribution of the model developed in this section is that it makes the linkage between *retrospective* responsibility and *prospective* responsibility (or criticality) concrete by providing a framework which relates these concepts in a quantitative manner and thus clarifies the mostly qualitative analyses provided in previous research (Berkowitz, 1972; Cane, 2002; Hart, 2008; Kerr, 1995; Schlenker et al., 1994; Weiner, 1995). Figure 4.23 shows that, generally, responsibility attributions increased with pivotality

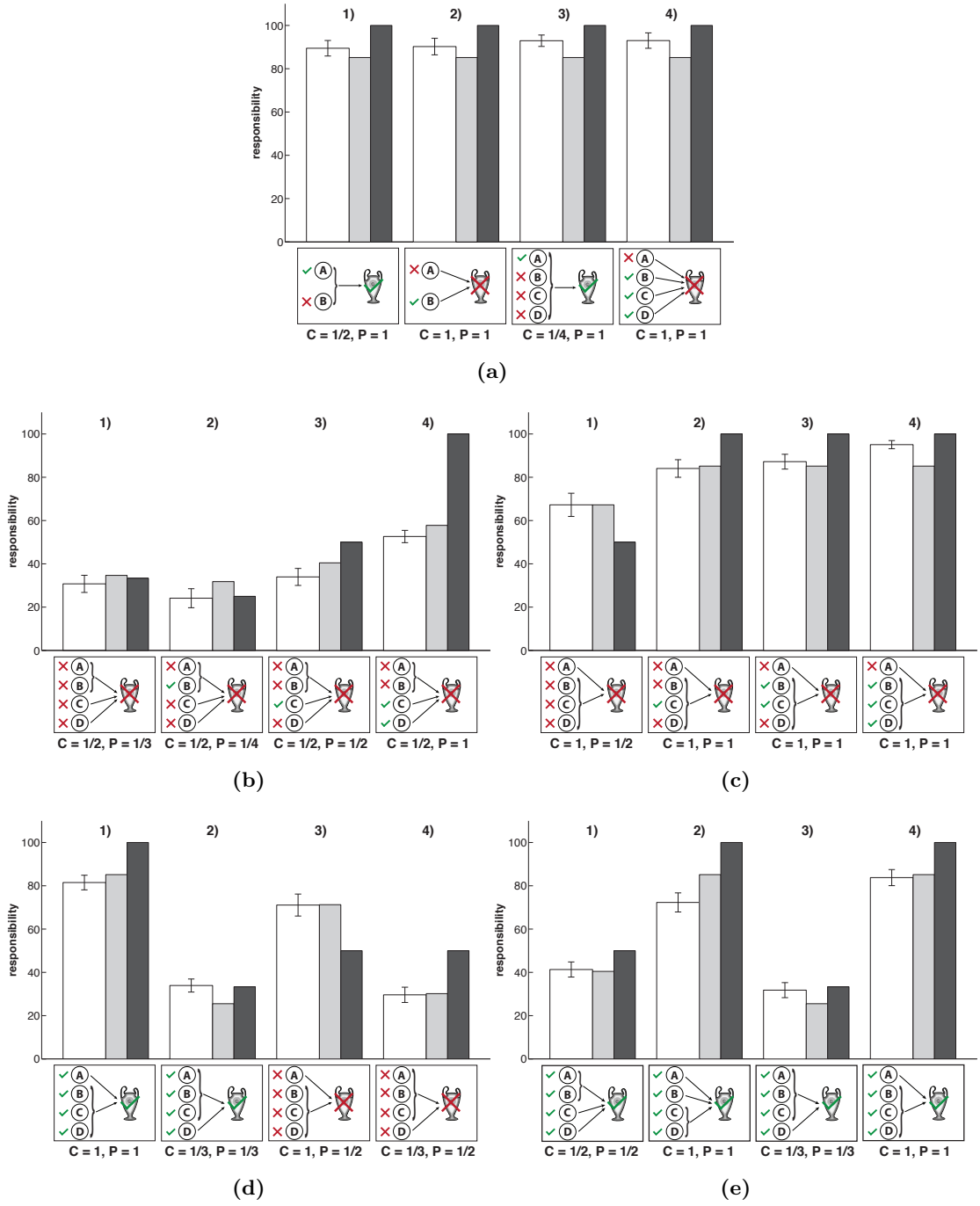
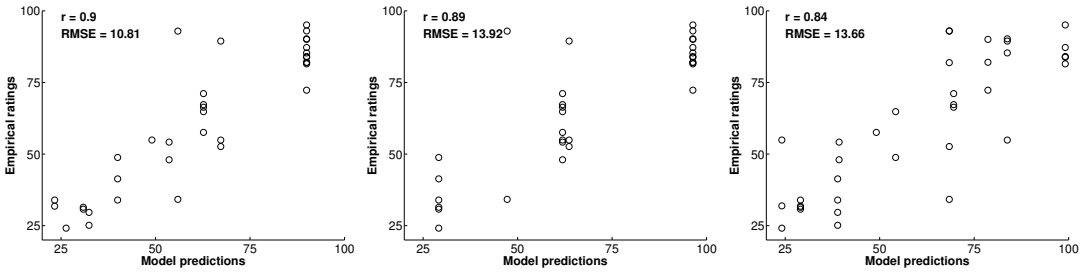


Figure 4.21: Mean responsibility ratings (white bars) for the remaining five sets of challenges with the predictions of the *critical-pivotality model* (gray bars) and the $\rho_{structural}$ model (black bars). Note: C = Criticality, P = Pivotality. Error bars indicate 1 SE.

(Figure 4.23a) and criticality (Figure 4.23b).[24]

Our approach is novel in that we propose a framework which combines concrete

[24]The finding that players who were 1/4 critical were overall more responsible than player who were 1/3 critical is due to the fact that in our experiment, players with a criticality of 1/4 happened to be overall more pivotal than players with a criticality of 1/3 (see Figure 4.23c).

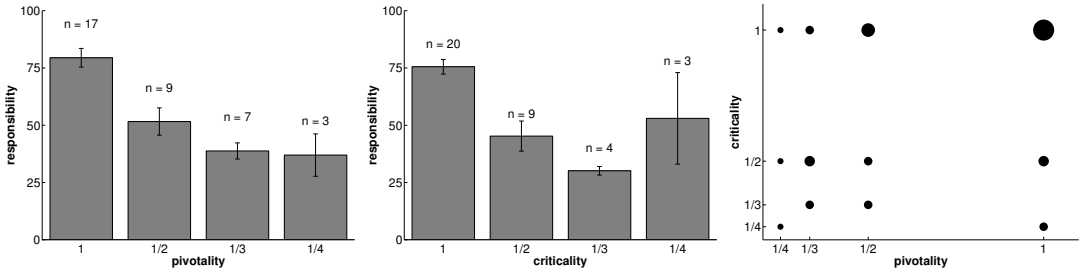


(a) $\pi_{heuristic} \times \rho_{structural}$ **(b)** $\pi_{heuristic} \times \rho_{counterfactual}$ **(c)** $\pi_{pivotal} \times \rho_{structural}$

Figure 4.22: Scatter plots for three implementations of the criticality-pivotality framework.

models of criticality and pivotality which allow us to derive quantitative predictions about how people will attribute responsibility. Although the current implementation in terms of a weighted mixture model has some degree of flexibility, it is nevertheless easily falsifiable. For example, the framework cannot predict that out of two players who are equally critical, the player whose contribution was closer to being pivotal is held *less* responsible for their team's outcome. Similarly, for two players who are equidistant from being pivotal, the model cannot predict that the less critical player receives *more* responsibility. Note, however, that none of these patterns of responses were observed in the 36 challenges we have used. Hence, thus far, the general framework stands.

The new model can also be applied to the data from Zultan et al.'s (2012) experiments. In fact, the new model makes the same predictions as the multiple path model proposed by Zultan et al. (2012) for the situations that we have tested in these experiments (compare, e.g., Figure 4.21b and 4.9). For more critical players, there are generally more ways to render them pivotal than for less critical players. Remember,



(a) Responsibility as a function of pivotality **(b)** Responsibility as a function of criticality **(c)** Correlation between pivotality and criticality

Figure 4.23: Mean responsibility attributions as a function of (a) pivotality and (b) criticality (n denotes the number of challenges for a given pivotality or criticality value), error bars indicate ± 1 *SE*. (c) shows the correlation between pivotality and criticality in our experiment, larger dots indicate more co-occurences (e.g. there were $n = 12$ challenges in which player A's pivotality and criticality was 1 but only $n = 1$ challenge in which A's pivotality and criticality was 1/4).

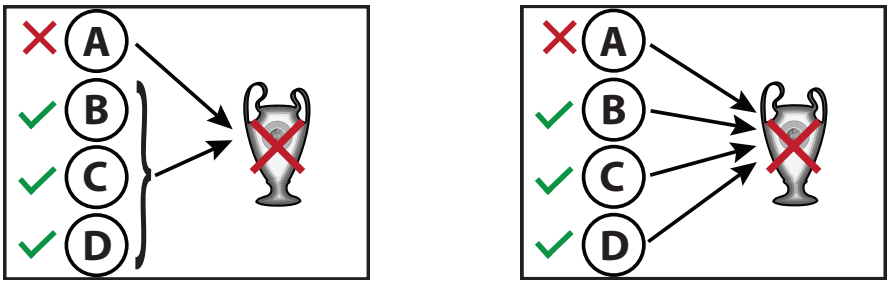
however, that the predictions of the two models are markedly different for players who are already pivotal. Whereas the multiple path model predicts that a pivotal player is always fully responsible, the criticality-pivotality model predicts that how responsible a pivotal player is seen depends on the degree to which that player's contribution was perceived to be critical for the group's outcome. The data from the last experiment hence support the new model and show that neither the structural model (Chockler & Halpern, 2004) nor the extended multiple path model (Zultan et al., 2012) can account for people's differential attributions to pivotal players.

Changes away from pivotality While the $\pi_{heuristic} \times \rho_{structural}$ model does a very good job in predicting participants' attributions overall, there is still some variance in people's responses that is left unaccounted for. In particular, participants attributed different degrees of responsibility to players who were fully critical and pivotal (see Figure 4.22a). One way of accounting for this variance is by extending the structural model of pivotality. In its current form, the model is only sensitive to how many changes are required in order to make a person pivotal. However, as discussed in the context of the triangle game results (Gerstenberg & Lagnado, 2010), maybe participants' attributions are not only affected by the number of changes *to* pivotality but also by the number of changes *away from* pivotality. In other words, it could be that people's attributions are influenced by how robustly a person was pivotal. If it would require many changes in order to render a pivotal person non-pivotal, then this person might be held more responsible for the outcome than another person for whom a smaller number of changes would suffice.

In the philosophical literature on causation, Woodward (2006) has made a similar claim for judgments of actual causation. He hypothesised that people are more willing to call one event C the actual cause of another event E when the counterfactual dependence relationship between C and E is insensitive to variations in the background conditions. Accordingly, both the sensitivity of the actual part $C \rightarrow E$ and the counterfactual part $\neg C \rightarrow \neg E$ matter. Thus, people feel comfortable to call C the cause of E if (i) C brings about E in a reliable fashion and (ii) E would not have happened if C had not happened even if the background conditions had been somewhat different. Applied to our example, when only a single change is sufficient to disrupt the counterfactual dependence between a player's action and the outcome, then that player's responsibility for the outcome is lower than when more changes would be necessary to undo the player's pivotal status.

Consider the two challenges depicted in Figure 4.24. In both situations, player A is fully critical and pivotal. However, there is a strong intuition that A is more responsible for the team loss in Challenge 1 compared to Challenge 2. One way to capture the asymmetry between these two situations is in terms of the notion of *changes away* from pivotality. While one change is sufficient to render player A non-pivotal in Challenge 2, three changes are required to make A non-pivotal in Challenge 1. Since more changes

are required to make A non-pivotal in Challenge 1 compared to Challenge 2, A is seen as more responsible in Challenge 1 compared to Challenge 2.



(a) Challenge 1 **(b)** Challenge 2

Figure 4.24: Two challenges in which player A is fully critical and pivotal.

Indeed, the data from our experiment already lend support to the assumption that people's responsibility attributions are not only sensitive to changes towards pivotality but also to changes away from pivotality. Consider, for example, the set of challenges shown in Figure 4.21c. A's criticality and pivotality is the same in Situations 2–4. However, participants blamed A significantly more for the loss in Situation 4 ($M = 95.08$, $SD = 11.89$) in which three changes away from pivotality were required compared to, for example, Situation 2 in which only one change would have rendered A non-pivotal ($M = 84.05$, $SD = 21.6$), $t(39), = 2.94, p = .006, r = 0.18$. Consider also the set of challenges shown in Figure 10d and compare the credit that A received in Situation 2 versus Situation 4. A is fully critical and pivotal in both situations. However, participants attribute significantly more credit to A in Situation 4 ($M = 83.78$, $SD = 23.44$) compared to Situation 2 ($M = 72.30$, $SD = 27.87$), $t(39), = 4.75, p < .001, r = 0.37$. This effect can also be explained by using the notion of robust pivotality. Whereas one change would be sufficient to render A non-pivotal in Situation 2 (i.e. changing B), three changes are required to render A non-pivotal in Situation 4. In contrast, in situations in which there are an equal number of minimal changes to render A non-pivotal (e.g. Figure 4.21a, Challenge 2 and 4), A's responsibility is the same. In future research, we will explore how both the distance from pivotality as well as the distance from non-pivotality affect responsibility judgments.

4.4 General Discussion

The experimental games we have used in this chapter to test the structural model (Chockler & Halpern, 2004) and to develop its extensions abstract away from many features of group interactions in everyday life. However, we see this as a feature rather than a limitation of our approach. By ruling out many possible confounds (e.g. differences in the underlying task or expectations about player's performance) we manage to

establish the importance of causality and counterfactual considerations for attributions of responsibility. By using experimental games, rather than scenarios, we were able to test participants' attributions for a host of different situations that differed in the underlying causal structure and the performance patterns. Furthermore, our modelling framework makes precise quantitative predictions about the degree to which each player will be held responsible in these different situations. Rather than only using labels for the core concepts in our framework, we have specified pivotality and criticality in terms of precise formal models.

This chapter has shown how this quantitative precision leads to refinements of existing theories (from Gerstenberg & Lagnado, 2010, to Zultan et al., 2012) and indeed novel theoretical developments (from Zultan et al., 2012, to Lagnado et al., accepted). Through testing the structural model and finding that people's attributions are not solely determined by pivotality, we were led to incorporate criticality as another crucial component into the model. The combined model explains the full pattern of results in our last experiment which contained 36 challenges with different structures and performance patterns. Now that the general modelling framework is set up, we can widen the experimental approach in several different ways and explore novel avenues. We are confident though that causality and counterfactuals, as incorporated in our framework, will remain as central pillars of a more full-blown computational model of responsibility.

In our experiments, we took great care to make sure that the individuals within the group always performed the same task such that differences in performance could not be explained away by assumed differences in the difficulty of the tasks.[25] By experimentally controlling for this factor, we could establish that differences in attributions were only due to manipulations of the group structure and performance patterns. However, in everyday life, groups often work together by sharing different tasks. In a football game, the task of a goalkeeper is different from the task of a striker. Similarly, when people work together in a team, their individual tasks are often quite different. For example, the joint efforts of a photographer, a journalist, an editor and many more combine in order to produce a newspaper article. A minimal way of rendering our general team setup more realistic would be by varying the difficulty of each players task. We could then ask questions such as whether a critical player is not blamed as much when he fails in a difficult as compared to an easy task (see Feather & Simon, 1971).

Not only does the difficulty of tasks within a group often vary but also the skill levels of the individuals within a group (Weiner & Kukla, 1970). Of course, difficulty and skill are closely related in that a difficult task will be comparatively easy for a highly skilled person. Both difficulty and skill have implications for the concepts of criticality and pivotality in our framework. We have already seen above that participants are sensitive

[25]This was not strictly the case for the cooking scenario in Zultan et al. (2012). However, even here, participants were instructed that the chefs had been randomly assigned to the different tasks and no indication was given that one task should be expected to be more difficult than the other.

to how the skill level of their partner affects how critical their own performance is for the group outcome (Kerr & Bruun, 1983). A highly skilled partner in a disjunctive task renders one's performance less critical. In contrast, if the task is conjunctive and the worst performance determines the team outcome then a skilled partner makes one's performance more critical. The opposite holds for a partner with low skill: it makes one's performance more critical in disjunctive and less critical in conjunctive tasks.

Formally, a person's skill level (or the difficulty of the task) can be expressed as the prior probability that a person will be successful in their task. B's prior affects how likely A will make a difference to the outcome (and thus how critical A is). A person's prior not only affects criticality but presumably also pivotality. Thus far, we have defined pivotality in terms of the minimal number of variables whose values need to be changed in order to make a person pivotal. We have already briefly discussed in the context of the triangle game that this notion of a minimal change is problematic when variables are not binary. We have seen that participants had a tendency to change the values of several variables slightly rather than making a single big change to one variable.

Even for binary variables, what counts as a minimal change is less clear when variables have different priors. Imagine that two players, A and B, failed in their individual tasks in a conjunctive team challenge. A's prior was high (A is more skilled) and B's prior was low (B is less skilled). For both players, one change in terms of variables is necessary in order to render them pivotal. However, intuitively, changing A from having failed to succeeded is a more minimal change than changing B. A was a priori likely to succeed and might have failed only due to some unstable factor which was present at that particular occasion (e.g. A's concentration might have been low). However, B in contrast, fails most of the time and succeeds only in very rare circumstances when all other conditions are perfect. Thus, it is more difficult to make A pivotal compared to B.

Varying the prior also highlights another problem: attributions of responsibility might not only be affected by how 'difficult' it is to render a person pivotal but also by how difficult it is to imagine that the pivotal person's performance would have been different. Thus, while it was easier to make B pivotal in the example above, changing the value of B (once she is pivotal) in order to change the outcome is more difficult than it is for A. In sum, we see that variations of the players' skill levels or the difficulty of their respective tasks lead to important questions about the notion of pivotality and what counts as a minimal change. Chapter 5 will report an experiment in which we explicitly varied the skills of group members and investigated how differences in skill affect responsibility attributions. We will also see different ways of how priors could be integrated into a model of responsibility attribution.

Of course, probabilities could also be introduced at the level of the causal structure. For example, it might be that the success of an individual within the group is only some-

times necessary in order for the team to be successful. This suggests that changes might not only be conceived of in terms of the behaviour of individuals but also in terms of the structure of the task. If the task had been slightly different, we would have won the game. More generally, it will be a fruitful avenue for future research to explore how normative considerations will influence responsibility attributions, whether varied through performance expectations or via what is morally appropriate in a particular situation (see Halpern & Hitchcock, forthcoming; Hitchcock & Knobe, 2009; Kahneman & Miller, 1986; Knobe & Fraser, 2008; McCoy et al., 2012; Sytsma et al., 2012).

Another way in which our current setup can be extended to further explore the notions of criticality and pivotality is by having more complex group structures (cf. Forsyth et al., 2002; Hamilton, 1978, 1986; Sanders et al., 2006). Arguably, the group structures we have used are very minimal: there are no direct interactions between the individuals within the group and their contributions combine via conjunctive (AND) and disjunctive (OR) functions. Most groups, however, have a much richer causal structure with mutual dependencies between the individuals. Our minimal setup allowed us to develop basic notions of criticality and pivotality. However, these notions will likely need to be revised in non-trivial ways once the group structures become more complex. Consider the three structures shown in Figure 4.25. In all three structures, the individual contributions of A, B and C combine in a conjunctive fashion (i.e. $P = min(A, B, C)$). Let us further assume, that the value of each variable is positive in all structures. How responsible is A for bringing about P in each structure?

In Structure 1, the application of our concepts of criticality and pivotality are straightforward. The criticality heuristic assigns full criticality to A and $N = 2$ changes are necessary to make A pivotal. However, things already become more problematic in Structure 2. In this structure, there is a fourth agent S who causes B and C. S could, for example, be a player whose successful performance is a precondition for B and C. Now how many changes are required in order to make A pivotal. One, two or three? The answer depends on how the notion of a change is operationalised. Are changes to be understood as interventions (see Pearl, 2000) or just as changes to the values of

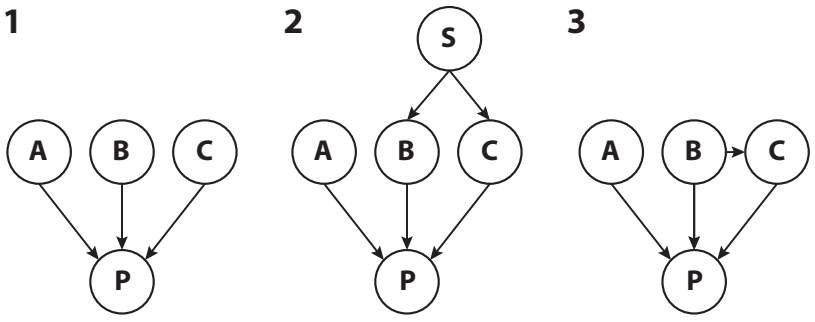


Figure 4.25: Causal structures with direct dependencies between agents.

the variables? If changes are interventions, then $N = 1$ change is sufficient to render A pivotal in Structure 2. If we intervene on S and make him fail his task, then B and C, the children of S, will also change their values. Causally dependent downstream variables are affected by the intervention. If, however, changes are not to be interpreted as interventions but in terms of the number of variables whose value needs to be different, then the answer would be $N = 2$ as the value of both B and C need to be different in order to make A pivotal. In Structure 3, there is a direct dependence between B and C in that B causally influences C. Again, this dependence has an influence on pivotality but also on criticality. Intuitively, B is more critical than both A and C. We are currently investigating responsibility attributions in richer causal structures which include hierarchical relationships between players.

Group structures can also be made more complex by considering alternative ways in which individual contributions combine to determine the group outcome. For example, imagine that a very strict bouncer at a night club in Berlin only grants entrance to couples or to groups who have an equal number of males and females. Hence, the group would only be successful and get in the club if the genders match. Alternatively, one can imagine situations in which a team only wins when an odd number of individuals are successful or when exactly a certain number of individuals succeed.

While we have used achievement-related tasks in order to explore our model, it will also be interesting to investigate how the model fares for situations of strategic interaction (such as the public goods games mentioned above, Rapoport, 1987) in which the variables represent decisions of different agents rather than their individual performances. A strategic setup could also be explored by putting participants in the role of a manager who has to assign different people with varying degrees of skill to different tasks. It would be interesting to investigate whether our principles of pivotality and criticality can explain how managers assign different people to different tasks. Our work suggests that there is a trade-off: on the one hand, a manager would like to make sure that the people in a team are maximally motivated (which is true for conjunctive tasks in which individuals perceive themselves to be highly critical). On the other hand, a manager has to make sure that the job is successful and thus introduce redundancies (i.e. disjunctive structures) which are potentially demotivating.

All these settings are in the general realm of what we can address with our framework. Although it might well be that our notions of criticality and pivotality will need to be refined in order to capture people's attributions, we are confident that the general principles will survive: people will be held more responsible to the extent that their contributions are perceived critical and that they were (close to) making a difference to the outcome.

The approach, as outlined at the moment, is very general. Indeed, responsibility is only defined in terms of criticality and pivotality whereby both notions are independent of the nature of the variables in the model. For example, to the model, it does not matter

whether a variable describes a physical event or an intentional agent. While this might be seen as inherently flawed, we think that it is a important feature of the framework. In the experiments we have reported, we did not vary people's epistemic states nor their intentions. It might well be that the results of our experiments would not have looked very different if we had replaced the participants in the game shows with robots or machines that need certain components to operate in order to function as a whole (see Halpern & Hitchcock, 2011; Hitchcock & Knobe, 2009). Our framework gives us the conceptual tools to tell what is required in order to get differential attributions between people and physical events. Chapter 2 has discussed a multitude of factors that have been shown to influence attributions of responsibility such as foresight and intentions. In future research, we will aim to incorporate these notions into our general framework to show how responsibility attributions differ between people and objects. This will also help us to potentially draw firmer distinctions between the related concepts of causality, responsibility and blame. Several experiments in Chapter 5 will corroborate that intentions and epistemic states are crucial for attributions.

While we have used groups with multiple players in order to generate differences in criticality and pivotality of individuals, these differences can also be investigated in situations in which only single agents are involved. As illustrated at the beginning of this chapter, situations of overdetermination do not require multiple agents. We might, for example, also ask whether the fact that John was just about to be hit by a massive boulder shortly after he was shot by Bill, reduces Bill's responsibility for John's death. In the discussion of the attribution literature in Chapter 2, we have seen that there are often several causes (some internal and others external) that jointly cause a particular action (e.g. luck, effort, ability and task difficulty in achievement contexts, see Weiner, 1995). Our framework can be used to explore how causality is divided between multiple factors in such situations (Kun & Weiner, 1973; Leddo, Abelson, & Gross, 1984; McClure, 1998; Reeder & Brewer, 1979).

Most of our experiments have focussed on how observers external to the group make attributions. As argued at the beginning of this chapter, having the observer external to the group is methodologically attractive as we retain full control over the stimuli that people see. If participants were part of the group, they would actively influence the experiment which would make comparisons between groups more difficult. However, future research should investigate to what extent the principles we have established for external observers also hold when people form part of the group. We have seen in Chapter 2 that people have a tendency for biased attributions of responsibility that foster their positive self-image. It might also be that people's sensitivity to causal structure, as established through our experiments, would influence the way in which people assign themselves to different available tasks. For example, it could well be that people seek tasks which maximise their chances of receiving credit when things go well and/or minimise their chances to be blamed when things go badly.

So far, we have not discussed any individual differences between participants in our experiments. However, of course, there were sometimes quite substantial differences in how participants attributed responsibility. Our framework incorporates several factors that are worth exploring as potential drivers of individual differences. For example, we might speculate that people differ in the extent to which they attribute responsibility to non-pivotal players. Whereas some might think that a person is only responsible for an outcome if her action could have made a difference, others might reason that a person is also responsible when the outcome is overdetermined.

People could also differ in how critical they perceive different players to be for the team outcome or in how much weight they give to criticality and pivotality. Although we offer objective definitions of pivotality and criticality, it is clear that people's attributions are influenced by *their subjective* estimation of these factors. Hence, two participants who have a different understanding of how a certain team structure works are predicted to arrive at different attributions of responsibility. The same player in the team might look very critical to one observer but not to another. This would particularly be the case when there was some uncertainty about the real underlying causal structure. If Peter, who went to a baseball game for the first time, and Paul, who is a baseball expert, were asked after a game how much different players were responsible for the outcome, they would likely reach very different conclusions. Peter's understanding of the causal relationships between the players and the (potential) impact of different actions is limited. Our framework suggests that arguments about to what degree someone is responsible for an outcome often boil down to arguments about the causal structure of the situation (or the rules of the game). More generally, our framework highlights the importance of investigating more thoroughly how people construct a mental causal model from the information they are provided with (see Fenton, Neil, & Lagnado, 2012).

Finally, let us briefly discuss whether our model is better described as a descriptive or a normative model of responsibility attribution. The answer is both. We developed our model as a psychological model with the aim to explain how people attribute responsibility to individual events when multiple causes are present. Indeed, our experiments have shown that the model does a good job in predicting people's attributions. We have also seen in Chapter 2 that people's attributions of responsibility are often motivated by their own interests. In most of our experiments, participants acted as external observers whose attributions did not have direct consequences on themselves. Hence, there was no incentive to attribute responsibility in any biased fashion. The fact that our model fares well descriptively and the finding that people are sometimes prone to attributional biases should not lead us to conclude that our model must be normatively inadequate. Indeed, we think that the two core factors, criticality and pivotality, do carry normative force. The degree to which a person is held responsible should be related to both whether her contribution was likely to make a difference *a priori* and whether it turned out to do so *ex post*.

A related question is whether our model is better understood as a *process model* or an *as-if model*. Whereas a *process model* makes explicit commitments about the underlying cognitive processes that are assumed to generate the observed behaviour, an *as-if model* is agnostic towards the cognitive mechanisms that give rise to the data. Again, we think that our model is a bit of both. On the one hand, we are convinced that people do consider how important a person's contribution is and that they might do so via simple heuristics such as the one we have outlined above. In more complex situations, people will have to think more deeply and maybe consider the chances of the team winning with or without the contribution of a player of interest in order to determine their criticality. Similarly, we think that counterfactual considerations are at the core of people's attributions of responsibility (see also Kahneman & Tversky, 1982).

On the other hand, we would not go as far as to assume that people actually go through the exact mental arithmetic that would be required to determine the minimal number of changes that are necessary to render each player pivotal. People might use more heuristic, distributive rules that take asymmetries between players into account and that approximate the predictions of our more sophisticated model. That being said, the fact that participants' patterns are so closely in line with pivotality thinking renders us confident that people do consider different possible situations that could have happened when attributing responsibility. We will see more direct evidence for counterfactuals in causal attributions in Chapter 6.

4.5 Conclusion

In this chapter, we have seen that there is a close relationship between causality, responsibility and counterfactuals. People's responsibility attributions to individuals within a team are sensitive to the causal structure of the group task which determines how the individual contributions combine to yield a group outcome. A structural model of responsibility (Chockler & Halpern, 2004) predicts people's attributions quite accurately in many situations. It correctly predicts that individual responsibility reduces when an outcome is overdetermined (Gerstenberg & Lagnado, 2010). Unlike a simple counterfactual model which predicts that people will only assign responsibility when a person is pivotal, people attribute responsibility to non-pivotal players, too. The structural model also correctly predicts that responsibility does not simply diffuse equally amongst the people in a group with an asymmetric structure but that it depends on the relationships between the individuals in the group (Zultan et al., 2012).

Finally, we have also seen that attributions of responsibility are not only influenced by how close a person was to being pivotal (Lagnado et al., accepted). It also matters how critical a player is perceived for the group's success a priori. Hence, two players who are both pivotal are not automatically attributed the same amount of responsibility. The player whose contribution was seen as more critical will be held more responsible. In-

deed, criticality can sometimes weigh stronger than pivotality: a critical but non-pivotal player is sometimes held more responsible than a less critical player who happened to be pivotal.

Chapter 5

Mental States and Responsibility

> **Captain Anabelle Brumford**: "I would now like to introduce the most distinguished American. This week he has been honoured for his 1000^{th} drug dealer killed. Ladies and Gentleman please welcome lieutenant Frank Drebin of police squad."
>
> **Frank Drebin**: "In all honesty, the last two I backed over with my car. Luckily, they turned out to be drug dealers."
>
> – **Excerpt from "The Naked Gun 2 1/2"**

In the previous chapter we have seen how people's responsibility attributions to individuals within a group are influenced by the ways in which the individual contributions combine in order to determine the group outcome. In all the experiments reported thus far, we have assumed that the mental states of the agents in the group are essentially identical: they all intend to win and try their best to achieve this goal. We also made sure that everything was common knowledge, that is, each player knew about the rules of the game and the position that she had randomly been assigned to in the team. Furthermore, we did not give participants any reasons to think that group members differed in their underlying skill levels. However, groups in everyday life are often much more diverse: they are comprised of individuals with different capabilities, knowledge is often unevenly distributed and individual intentions are seldom completely aligned. In Chapter 2, we have seen that mental states, such as foresight and intention, feature prominently in theories of responsibility attribution. In this chapter we will explore how these factors influence responsibility attributions as well as decisions about economic interactions. The remainder of the chapter is organised as follows.

In the first section, we will see how differences in agents' epistemic states affect people's attributions of responsibility. In the second section, we will look at how differences in individual skill level affect responsibility attributions. In the third section, I will focus

on the role of intentions for attributions of responsibility and for how people interact in an economic game. In concluding, I will discuss how the two major aspects of responsibility attributions in groups, causal structure (Chapter 4) and mental states can be studied in unison. I will highlight the need for a modelling framework in which both structural aspects as well as the mental states of the agents can be incorporated.

5.1 Knowledge (Gerstenberg & Lagnado, 2012)

In this section, we will investigate how group members' knowledge about each other's contributions in a causal chain affect attributions of responsibility as a function of whether or not the group outcome is overdetermined. Imagine that you are the coach of your country's relay team. In what order would you make your runners compete? Would you put the best runner first, last, or in one of the middle positions? How responsible would the runners in the different positions be if the team won or lost?

According to a simple counterfactual analysis, each of the individual events in a causal chain qualifies equally as a cause of the final effect. If any of the events in the chain had not occurred, the effect would also not have occurred. However, several studies have shown systematic differences as to which events in a chain are judged to be more causal (Miller & Gunasegaram, 1990) or more likely to be mentally undone in order to prevent the outcome from happening (Wells et al., 1987).

In Chapter 2, we have introduced Spellman's (1997) *crediting causality model* (CCM), which predicts that an event's perceived causal contribution varies with the extent to which it changes the probability of the eventual outcome.[1] The more an event changes the outcome's probability, the more it is judged to be causal. Accordingly, the model predicts a *primacy effect* when the first event in a causal chain changes the probability of the outcome more than any of the later events. Conversely, it predicts a *recency effect* when the probability change is greatest for the final event in a chain.

Since its proposal, several shortcomings of the CCM have been demonstrated. Importantly, because the model predicts causality ratings merely on the basis of the statistical notion of probability change, it is insensitive to the ways in which these changes are brought about. However, studies have shown that voluntary human actions are preferred over physical events as causes (Lagnado & Channon, 2008), even when the changes in probability are identical (Hilton et al., 2010). Furthermore, Mandel (2003) showed that a later event can receive a higher causal rating, even though an earlier event has already increased the probability of the outcome almost to certainty. If a victim

[1] Although Spellman's (1997) model was originally developed for judgments of causal contributions, it was also used to predict judgments about blame and experienced guilt. In our experiments, we used the term blame for negative responsibility and credit for positive responsibility, to highlight the outcome-dependent valences of responsibility attributions. However, we acknowledge that these terms are not equivalent and that situations exist in which the results will be affected by choice of terminology (see, e.g., Robbennolt, 2000).

has been poisoned first but is then killed in a car crash, people select the car crash as the cause of death rather than the poison, despite the fact that the poison had already increased the probability of death to certainty.

In this section, we highlight a different problem that has not been addressed by previous research and motivate a theoretical extension of the model. The CCM predicts that people's attributions of responsibility are determined by comparing what actually happened with what would have happened had an event in *this particular situation* been different. However, we argue that attributions of responsibility are affected not only by the degree to which an event made a difference in the particular situation in which it occurred, but also by what would have happened had the event of interest been different in *other similar situations* (see Chockler & Halpern, 2004, for a formal model that incorporates this idea, and Gerstenberg & Lagnado, 2010; Lagnado et al., accepted; Zultan et al., 2012, for empirical support).

Consider the example of a team relay, mentioned earlier. If the performances of the first three runners in a team were very poor, the probability of the team winning before the fourth runner started would be essentially zero and could not be increased any more, irrespective of the fourth runner's performance. Since the last runner's performance did not make a difference to the team outcome in this particular situation, the CCM would predict that this runner's responsibility would be low, independent of whether she performed well or poorly.

However, when we consider not only the actual situation, but also other possible situations, it becomes clear how responsibility attributions could still be sensitive to differences in performance. The athlete who performed well despite the certainty of the team's loss could have made a difference to the outcome *if* the other team members' performance had been better. In contrast, an athlete who performed poorly in the same situation would send an ambiguous signal: it could be that she did not try hard, because the team outcome was already determined. Yet it is also possible that the athlete would not have performed better even if a situation arose in which the final contribution was required. This difference in uncertainty over whether or not an athlete is capable of performing well licenses a differential attribution of responsibility. Despite the fact that the difference in performance did not matter in the actual situation, it would have made a difference in situations in which the performance of the other three team members had been better (cf. Reeder & Brewer, 1979).

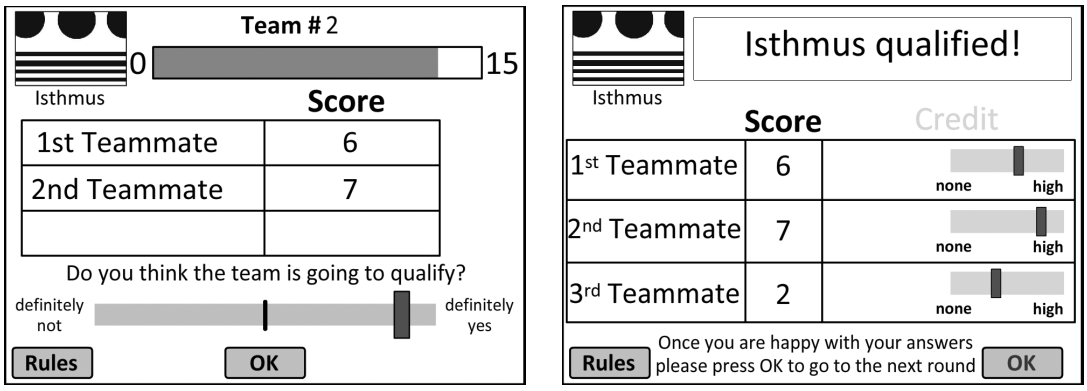
In the present study, we explored how (1) level of performance and (2) the extent to which a contribution was critical to the result, as measured by the change of the outcome's probability that the contribution induced, affect people's perceptions of how responsible each contribution is for the eventual outcome. In line with CCM, we expected that the extent to which identical performance would be seen as responsible for the team's result would vary depending on whether or not the result was already determined. We also predicted that responsibility attributions would be influenced not only by how

much a person's contribution made a difference in the actual situation, but also by whether it could have made a difference in other possible situations. Hence, we expected that the level of performance of an individual player would influence how responsible she was seen to be for the team's outcome, even in situations in which the result had already been determined prior to her performance.

5.1.1 Experiment 1

Participants acted as external observers evaluating the performance of different teams in the Olympic qualifiers of an invented sport. Each of 32 countries was represented by a team of three athletes. The athletes performed their routines individually and received a score from a panel of judges ranging from 0 (very bad performance) to 10 (excellent performance). Participants were instructed that the average performance in the competition was 5 points. A country would qualify for the Olympics if its team scored 15 points or more in total. It did not matter how many points a team got once they were above this qualifying standard. Participants were informed that the athletes performed their individual routines sequentially and that later athletes *knew* how their previous teammates had performed. For each of the 32 teams, participants experienced two different phases. In the *probability rating phase*, they saw the scores of each of the three athletes sequentially and, after each athlete's score, indicated on a slider how likely they thought it was that the team would qualify (see Figure 5.1a). The slider ranged from 0 ('definitely not') to 100 ('definitely yes') and was initialised at the midpoint. The progress bar at the top of the screen showed how many points were still required for the team to qualify and was updated after each athlete's score. Once the team qualified, the progress bar turned green. If the team could not qualify anymore, the bar turned red.

In the *responsibility attribution phase*, each athlete's score was shown simultaneously



(a) Probability updating phase **(b)** Responsibility attribution phase

Figure 5.1: Screenshots of the experiment.

Table 5.1: Patterns of athletes' scores in the experiments. *Note*: The team as a whole has to achieve a score of 15 or more in order to win.

Situation	Scores	Athlete 1	Athlete 2	Athlete 3 low score	Athlete 3 high score
certain loss	non-identical	1	1	4	8
		1	2	3	7
		3	1	2	6
	mean	1.67	1.33	3	7
	identical	1	1	1	
		2	2	2	
uncertain loss	non-identical	4	2	4	8
		2	5	3	7
		4	4	2	6
	mean	3.33	3.67	3	7
	identical	3	3	3	
		4	4	4	
uncertain win	non-identical	5	6	4	8
		7	5	3	7
		6	7	2	6
	mean	6	6	3	7
	identical	6	6		6
		7	7		7
certain win	non-identical	7	9	4	8
		9	8	3	7
		9	9	2	6
	mean	8.33	8.67	3	7
	identical	8	8		8
		9	9		9

mean = average score of each athlete for the patterns with non-identical scores

in a table (see Figure 5.1b). Participants were asked, "To what extent is each of the athletes responsible for their team's success or failure to qualify?" If the team qualified, participants attributed credit (green sliders ranging from the center to the right). If the team did not qualify, participants attributed blame (red sliders ranging from the center to the left). The sliders for each athlete ranged from 0 ('none') to 10 ('high') and could be moved independently; that is, they did not have to sum to a certain value.

Table 5.1 shows the patterns of scores that were used in the experiment. We systematically varied the scores of the third athlete to be either low or high for different scores by the first two athletes. This approach allowed us to compare how an identical

performance of the third athlete would be evaluated, as a function of whether the result was already certain prior to the final athlete's turn or was still uncertain. There were two possible ways in which the team's result could already have been determined by the scores of the first two athletes. A team's loss was certain if the sum of the first two athletes' scores was 4 points or less. Because the maximum score that an athlete could achieve in the challenge was 10, it was impossible in this case for the third athlete to allow the team to win. Likewise, a team's win was certain prior to the third athlete's performance if the first two athletes' scores added up to 15 or more points.

A consequence of this design was that, while we kept the absolute performance of the third athlete identical in the different situations, the *relative* performance as compared to the teammates varied. The final athlete performed relatively well in the *certain loss* as compared to the *uncertain loss* cases, and relatively poorly in the *certain win* as compared to the *uncertain win* situations. Because our main interest concerned the effect of the (un)certainty of the outcome on the attributions for the third athlete, we controlled for the effects of relative performance by including eight additional cases in which the scores of all three athletes were identical. Here, all athletes either scored 1 (or 2) in the *certain loss* cases, 3 (or 4) in the *uncertain loss* cases, 6 (or 7) in the *uncertain win* cases, and 8 (or 9) in the *certain win* cases. Any differences between the three athletes in these situations can only be explained in terms of order effects.

The main target of interest in our design was the third athlete. We hypothesised that both her performance and the certainty of the team's result prior to her turn would affect participants' responsibility attributions. In line with the CCM, we predicted that the third athlete would receive less credit for an identical performance if the result was already certain rather than still uncertain. Likewise, we predicted that the athlete would receive less blame if the team had already certainly missed the qualification threshold prior to her turn. However, in contrast to the CCM, we expected the third athlete's blame for losses to be higher and her credit for wins to be lower when she received a low rather than a high score, even in situations in which the results were already certain.

5.1.1.1 Methods

Participants A group of 41 (22 female, 19 male) participants recruited through the UCL subject pool took part in this experiment. Their mean age was 23.1 years (SD = 2.5).

Materials The program was written in Adobe Flash CS5.[2]

Design For the 24 patterns in which the scores of the three athletes were non-identical, the experiment followed a within-subjects 2 ('Result': win vs. loss) × 2 ('Certainty of

[2]The experiment can be accessed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/order_demo.html`

Outcome': uncertain vs. certain) × 2 ('Performance of the Third Athlete': low vs. high score) design. For the eight patterns in which the scores of the three athletes were identical, the experiment followed a within-subjects 2 ('Result') × 2 ('Certainty of Outcome') design.

Procedure The study was carried out online. After having read the instructions, participants did one practice trial in which the different components of the screen were explained. A set of four comprehension check questions ensured that participants had understood the task. On average, they answered 89% of the comprehension check questions correctly. After they had answered each of the questions, the correct solution was displayed. Participants then evaluated the performance of 32 teams in the probability rating phase (Figure 5.1a) and the attribution phase (Figure 5.1b), as described above. If a team did not qualify, participants attributed blame; otherwise, they attributed credit. Throughout the experiment, they could remind themselves of the rules by clicking on the 'Rules' button at the bottom left corner of the screen. The median time that it took participants to finish the study was 18.8 minutes.

5.1.1.2 Results

Probability updating phase Figure 5.2 shows the mean probability-of-success ratings for wins and losses, separated for situations in which the outcome was either already certain after the second athlete's score or still uncertain. The results of the probability updating phase did not differ between Experiments 1 and 2; hence, we report here the

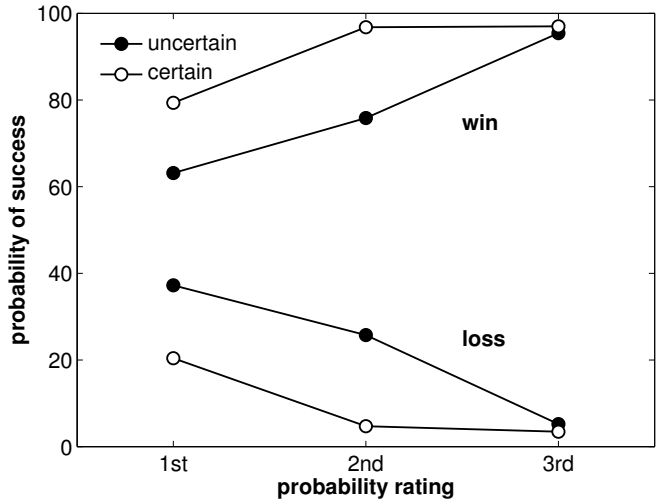


Figure 5.2: Mean rated probabilities of success for wins (top) and losses (bottom) as a function of whether the outcome was uncertain (black) or certain (white) after the second player's score. The data are aggregated across Experiments 1 and 2.

aggregated data of both experiments.[3]

For losses, participants' probability-of-success ratings after the second athlete's score was revealed were significantly lower in the certain ($M = 4.7$, $SD = 7.6$) than in the uncertain cases ($M = 25.7$, $SD = 9.9$), $t(118) = -26.44, p < .05, r = .43$. For wins, participants' probability ratings were significantly higher in the certain ($M = 96.8$, $SD = 8$) than in the uncertain ($M = 75.8$, $SD = 9.85$) cases, $t(118) = -28.18, p < .05, r = .44$.

Responsibility attribution phase First, we wanted to test the extent to which the blame and credit ratings for the third athlete would vary as a function of performance and of whether or not the team's result was already certain after the second athlete's score. Two separate 2 (*certainty*: certain vs. uncertain) × 2 (*performance*: low vs. high score) repeated measures ANOVAs for losses and wins were conducted on the ratings for the third athlete in the non-identical cases (see Figure 5.3 top).

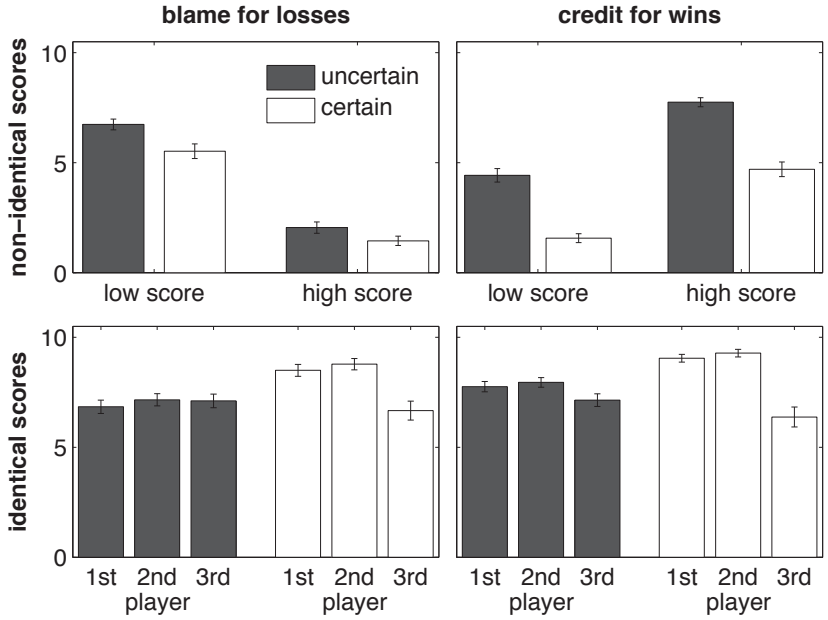


Figure 5.3: Mean blame and credit attributions for Experiment 1. The top panels show mean attributions for the third player in the nonidentical-score cases, and the bottom panels show mean attributions for all three players in the identical-score cases. Error bars indicate ±1 *SE*.

For losses, we found significant main effects of performance, $F(1, 40) = 129.33, p < .05, {\eta_p}^2 = .764$, and certainty, $F(1, 40) = 6.86, p < .05, {\eta_p}^2 = .146$, but no interaction. The third athlete was blamed more if she received a low rather than a high score. Furthermore, her blame ratings were lower when the team had already certainly missed the qualification criterion, as compared to when the outcome was still uncertain. Crucially,

[3]Experiments 1 and 2 were identical except for a manipulation in the instructions that did not affect the probability updating phase (see below).

the effect of performance significantly influenced the athlete's blame ratings for situations in which the outcome was still uncertain, $t(40) = 11.37, p < .05, r = .47$, as well as when the outcome was already determined, $t(40) = 8.36, p < .05, r = .42$.

For wins, we found significant main effects of performance, $F(1, 40) = 66.03, p < .05, {\eta_p}^2 = .623$, and of certainty, $F(1, 40) = 31.76, p < .05, {\eta_p}^2 = .443$, but no interaction effect. The third athlete received more credit for a high than for a low score. Also, credit attributions were higher if the result was still uncertain, as compared to when it was certain. Again, the effect of performance significantly influenced the athlete's credit ratings for situations in which the outcome was still uncertain, $t(40) = -7.76, p < .05, r = .40$, as well as when it was already determined, $t(40) = -6.84, p < .05, r = .38$. As outlined above, the predicted order effect and the relative performance effect go in the same direction for the cases in which the scores of the three athletes were nonidentical. For example, the third athlete performed relatively well as compared to the others in situations in which the loss was certain rather than uncertain after the second athlete's performance (see Table 1). Hence, we analysed the situations separately in which all of the athletes had identical scores. Any differences for these cases could only be explained with respect to the order of performance.

Figure 3a (bottom) shows the mean blame ratings for losses and credit ratings for wins attributed to all three athletes in the team in situations in which the result was either uncertain or certain. To evaluate whether the third athlete's ratings varied as a function of certainty of outcome, we compared the difference in the average attributions to the first two athletes with the attributions to the third athlete. For losses, this difference was significantly greater in the certain cases (M = -1.97, SD = 3.77) than in the uncertain cases (M = 0.11, SD = 1.81). The third athlete received significantly less blame for an identical performance if the result was already certain as compared to uncertain, $t(40) = -3.88, p < .05, r = .30$. For wins, likewise, the third athlete received significantly less credit for an identical performance if the result was already certain (M = -2.79, SD = 4.18) than if it was uncertain (M = -0.71, SD = 2.16), $t(40) = -3.29, p < .05, r = .28$.

5.1.1.3 Discussion

The results of Experiment 1 revealed an *attenuation effect*: blame and credit attributions to the last athlete were reduced when the result was certain rather than uncertain. However, how much blame (or credit) an athlete received for the team's loss (or win) also depended to a large extent on her performance. Importantly, even when the result of the team challenge was already determined, an athlete still received more credit and less blame for a good than for a bad performance.

5.1.2 Experiment 2

In Experiment 2, we investigated whether the attenuation effect observed in Experiment 1 for the situations in which the outcome was already determined would overgeneralise to situations in which it would arguably be inappropriate. As outlined above, an athlete who achieves a low score in a situation in *which she knows* that the outcome is already determined sends an ambiguous signal: she could have achieved a low score either for not having tried hard or for just not being capable of performing well. Crucially, an athlete can only justify not trying hard in a situation in which she knows that the team result is already certain. Our paradigm allowed us to dissociate the 'objective certainty' of the team result from the perspective of an external observer from the 'subjective certainty' of the team result from the perspective of an athlete in the team. For Experiment 2, we kept the experienced order of events unchanged but altered the knowledge states of the athletes. As in Experiment 1, participants experienced the scores of the three athletes sequentially; however, they were instructed that the athletes did *not* know about each other's scores.

Spellman (1997) has shown that attributions are influenced not only by the experienced order of events but also by the order in which the events occurred in the world. To test for any effects of the objective order of events, we manipulated whether athletes were described as competing simultaneously (Experiment 2a) or sequentially (Experiment 2b). Would participants show reduced blame attributions for losses and credit attributions for wins when they knew that the result was already determined? Or would they appreciate that the athletes did not know each other's scores, and hence would show no attenuation effect as a function of the certainty of the result?

5.1.2.1 Methods

Participants A group of 56 participants (42 female, 14 male) from the UCL subject pool participated in Experiment 2a and 22 participants (15 female, 7 male) recruited through Amazon Mechanical Turk took part in Experiment 2b. The mean age was 25.36 years ($SD = 9.8$).

Design and materials The design and materials were identical to those of Experiment 1. 'Objective Order of Events' (simultaneous vs. sequential) was included as between-subjects factor.

Procedure The procedure was identical to that of Experiment 1, except for a minimal change in the instructions: participants were informed that the athletes of each team performed their individual routines either simultaneously (Experiment 2a) or sequentially (Experiment 2b). Importantly, participants were instructed that the athletes *did not know* their teammates' scores. On average, participants answered 89% of the com-

prehension questions correctly. The median time that it took participants to complete the study was 16.5 minutes.

5.1.2.2 Results

Responsibility attribution phase None of the comparisons were significantly influenced by the between-subjects factor 'Objective Order of Events'. Hence, we combined the data of both conditions.

Again, the blame and credit ratings for the third athlete varied with how the athlete performed and, although to a much lesser extent, with whether or not the result was already certain. Figure 5.4 (top) shows the blame and credit attributions of the third athlete for the situations in which the scores of the three athletes were nonidentical. For losses, we found significant main effects of performance, $F(1, 77) = 433.08, p < .05, {\eta_p}^2 = .849$, and certainty, $F(1, 77) = 13.53, p < .05, {\eta_p}^2 = .149$, as well as an interaction effect, $F(1, 77) = 4.52, p < .05, {\eta_p}^2 = .55$. For wins, we found significant main effects of performance, $F(1, 77) = 544.32, p < .05, {\eta_p}^2 = .876$, and certainty, $F(1, 77) = 24.77, p < .05, {\eta_p}^2 = .243$, but no interaction effect.

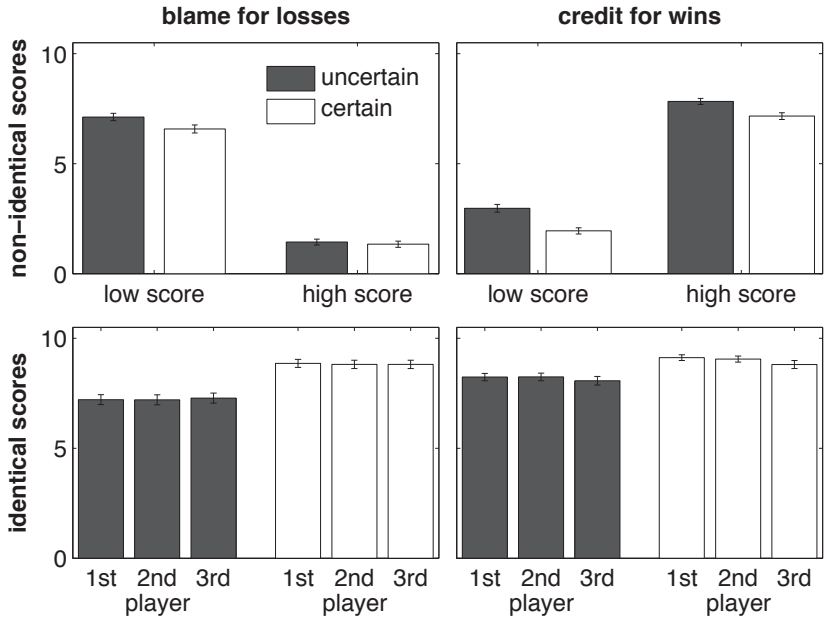


Figure 5.4: Mean blame and credit attributions for Experiment 2 (collapsed across both conditions). The top panels show mean attributions for the third player in the nonidentical-score cases, and the bottom panels show mean attributions for all three players in the identical-score cases. Error bars indicate ±1 *SE*.

Importantly, the difference between the blame attributions in the certain cases and the uncertain cases was significantly larger in Experiment 1 ($M = 0.91$, $SD = 2.23$) than in Experiment 2 ($M = 0.23$, $SD = 0.70$), $t(117) = 2.12, p < .05, d = 0.41$. Likewise, the differences between the credit attributions as a function of outcome certainty were

larger in Experiment 1 ($M = 2.95$, $SD = 3.35$) than in Experiment 2 ($M = 0.40$, $SD = 0.81$), $t(117) = 4.73, p < .05, d = 0.91$.

Figure 5.4 (bottom) shows the blame and credit ratings of all three athletes for the situations in which their scores were identical. There was no significant difference between the mean blame ratings for the first two athletes and the third athlete in either the certain cases ($M = 0.02$, $SD = 0.44$) or the uncertain cases ($M = -0.07$, $SD = 0.7$). Furthermore, there was also no significant difference for the credit ratings between the certain ($M = 0.28$, $SD = 1.09$) and the uncertain ($M = 0.17$, $SD = 0.68$) cases.

5.1.2.3 Discussion

As in Experiment 1, the blame and credit ratings varied depending on how the athletes performed. However, in contrast to the results of Experiment 1, the 'Certainty of Outcome' factor now had only a small influence on participants' responsibility attributions for the third athlete in the nonidentical cases, and no influence in the identical cases. As mentioned above, the small effect of the 'Certainty of Outcome' factor was likely a result of the differences in relative performance in the nonidentical cases. Most of the participants were not influenced by the experienced order of events and took into account the fact that the third athlete did not know her teammates' scores. Hence, the level of the third athlete's performance was not discounted in situations in which the outcome was already determined. In addition, whether the events actually happened simultaneously or sequentially did not affect attributions.

5.1.3 General discussion

The results of two experiments showed that people exhibit an attenuation effect in their responsibility attributions for team members contributing sequentially to a team challenge. As predicted by the CCM, the level of a team member's performance is discounted for situations in which the outcome is already determined. Importantly, this effect is only present in situations in which the later team member knows the previous members' scores, and it is not affected by the objective order of events. The effect does not overgeneralise to situations in which the order in which participants learn about the scores does not map onto the order in which they are generated: The attenuation effect is moderated by the inferred epistemic states of the athletes.

In both experiments, attributions of responsibility were also strongly affected by the level of performance: Athletes generally received more credit and less blame for a high score than for a low score. Importantly, the effect of performance was present even in situations in which the team result was already certain and known to the athletes. We take this finding as evidence that participants' responsibility attributions are not solely determined by whether an individual's contribution made a difference in the actual situation.

The CCM can be extended in a natural way to capture the patterns of responsibility attribution from our study. Rather than determining an individual's causal responsibility solely in terms of the difference her contribution made to the team outcome in the actual situation, it also matters whether her contribution was likely to have made a difference in other possible situations – namely, situations in which the team result was not yet certain prior to her contribution.

We do not think that participants' sensitivity to differences in performance in situations in which the result was already certain is unreasonable. The level of performance of the third athlete conveys important information. If, for example, an athlete performed well despite the fact that the team had already lost for sure, we would learn that the athlete was in principle capable of good performance. This good performance could have made a difference, in a counterfactual situation in which the teammates had performed somewhat better. In contrast, if the third athlete performed poorly, we could not be sure whether this was due only to the fact that she knew that the team had already lost, and hence did not try hard, or whether she could not have performed better even if the contribution had been needed.

Inferences about a person's underlying skill level based on their performance fit into what Reeder and Brewer (1979) have termed a *hierarchically restrictive schema* of behaviour-to-disposition inference. According to this schema, there is an asymmetrical mapping between disposition and behaviour. While a high standing on the dispositional attribute is consistent with a wide range of behaviours, lower values on the disposition are only consistent with a smaller range. Put differently, a person's standing on the dispositional continuum determines the upper bound of behaviours that can be expressed. A skilled athlete, for example, is capable of a very good performance but could also produce a poor performance due to lack of motivation. An unskilled athlete, in contrast, does not have the capability of producing a very good performance.[4] When reasoning diagnostically from an observed behaviour to the underlying disposition, the *hierarchically restrictive schema* implies that a very good performance is more diagnostic for the person's skill than a poor performance.

Furthermore, taking an individual's performance into account also makes sense if one considers the forward-looking motivational function that attributions of responsibility can serve (cf. Weiner, 1995): receiving blame for a negative outcome or credit for a positive outcome motivates better performance in the future. Despite the fact that how one performed might not have been critical in the present situation, a high performance might well make the crucial difference in a future situation.

Our theoretical extension of the CCM makes straightforward predictions. The degree to which a contribution is judged as being responsible for the outcome will not only increase as a function of whether it made a difference in the actual situation, but

[4]Unless, of course, the task involves a considerable degree of luck but even then, a consistently high performance would not be expected from an unskilled person.

also with an increased chance that it would have made a difference in similar possible situations. Conversely, responsibility is expected to be low when a contribution does not make a difference in the actual situation and is also unlikely to make a difference in other possible situations (cf. Wells & Gavanski, 1989)

Our proposal is in line with recent work that has highlighted the close relationship between counterfactuals and attributions of causality and responsibility (Chockler & Halpern, 2004; Halpern & Pearl, 2005; Petrocelli et al., 2011). As we have seen in Chapter 3, according to Chockler and Halpern's structural model of responsibility, an agent is maximally responsible if she made a difference in the actual situation. However, responsibility does not immediately drop to zero when an agent's contribution made no difference. Rather, an agent's responsibility decreases as a function of the distance of the actual situation from a situation in which her contribution would have made a difference. One of the important advantages of this approach is that it can deal with cases of causal overdetermination, in which there are multiple sufficient causes. While each individual cause actually makes no difference to the outcome in such a situation, there is a possible situation for each cause in which it would have made a difference – namely, the situation in which the other causes were absent. The further away such a situation is from the actual situation (i.e., the more changes would need to be made), the less responsible is the cause under consideration.

While the results of this experiment support the proposed extension of Spellman's (1997) model, alternative explanations cannot be ruled out on the basis of the data from the reported experiments alone. For example, one might argue that participants' responsibility attributions are mostly determined by the athletes' scores and that the attributions to the third athlete in the first experiment were subject to an anchoring-and-adjustment process. While such a performance-adjustment account could explain the results for the reported experiments, in which individual performance was measured by a discrete variable, such an account would have difficulty capturing situations in which performance was measured by binary variables. As we have seen in Chapter 4, Zultan et al. (2012) and Lagnado et al. (accepted) have shown that in situations in which each team member's performance is binary (i.e., they can either succeed or fail in their individual task), responsibility attributions for the team outcome are still sensitive to how close an individual's contribution was to making a difference.

In sum, our experiments highlight the fact that a comprehensive model of responsibility attribution in group contexts will need to take into account the individuals' mental states (Gerstenberg et al., 2010), the extent to which each contribution made a difference to the outcome in the actual situation, and whether the contribution could have made a difference if things had turned out somewhat differently (Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2010).

5.2 Expectations (Gerstenberg et al., 2011)

In this section, we will have a look at how manipulations of players' skill levels and associated differences in performance expectations influence responsibility attributions in an achievement context. Consider you witness the following situation on a night out in a pub. Three friends are playing darts and, to spice things up, one of them offers the other two the following deal: "Both of you throw at the same time. If one of you manages to hit the dart in the centre region, the next round of drinks will be on me. However, if neither of you hits the centre, you'll have to pay for my next pint."

You have seen from their previous play that one of the players is very skilled. In fact, she managed to hit the centre region most of the time. The other player's performance, in contrast, was quite poor. He hardly ever managed to get the dart in the centre. How would you spread the blame if neither of them managed to hit the centre? Whom would you credit more if both of them hit the centre? This paper investigates how people attribute responsibility between multiple agents based on their underlying skill levels and actual performances.

Skill, expectation and control The problem of how credit for a positive outcome or blame for a negative outcome should be distributed across the members of a group is encountered in many contexts – from law, business and medicine, to heated dinner table debates about team sports. Skill and performance are important variables that potentially differentiate the individual agents contributing to a joint effort and are hence likely to influence credit and blame attributions. How skilled we think a person is, has a direct influence on what performance we expect from her. Furthermore, skill is closely connected to the notion of control. If a person is skilled it implies she is able to do something well in a reliable fashion. However, as the well-known phenomenon of choking in sports demonstrates, a player might fail to deliver because he struggles with the external pressure imposed by high expectations. Hence, high skill does not necessarily imply good performance. Similarly, a low skilled person can sometimes surprise with a very good performance. How do considerations about skill and performance influence people's attributions of blame or credit and what cognitive processes are likely to guide responsibility attributions in these contexts?

Achievement motivation: Ability and effort We have seen in Chapter 2 that a rich literature in attribution research has been concerned with analysing the causal factors that are perceived to influence an agent's success or failure in achievement related contexts (see, e.g. Weiner, 1995). Weiner and Kukla (1970, Experiment 1) presented scenarios in which they systematically varied the ability and effort of hypothetical students paired with different performance outcomes. For example, a student could be described as having low ability, expended high effort and achieved an excellent grade in their exam. Based on this information, participants were asked to assign reward or

punishment to the students. The results showed that whether students received punishment or reward was directly related to the outcome of their exam whereby participants showed a tendency to reward more than punish. Additionally, participants' responses were significantly influenced by both the student's ability and effort. Students who expended high effort were rewarded more and punished less than students who expended low effort. Furthermore, students with high ability received more punishment and less reward compared to students with low ability. Interestingly, whereas both able and non-able students received the same reward for the best possible outcome (an excellent exam), able students received more punishment than non-able students for the worst possible outcome (a clear failure in the exam). Overall, however, reward and punishment were more strongly influenced by differences in the expended effort than ability.

In order to explain this difference, controllability has been identified as an important factor that distinguishes effort from ability (Weiner, 1995). As discussed in Chapter 2, Alicke's (2000) model of culpable control draws a useful distinction between *behaviour control* and *outcome control*. Whereas how much effort we expend is a *behaviour* we have control over, we cannot behaviourally control our ability at a given moment. How much causal control a person has over an *outcome*, however, depends to a large extent on the person's ability (as well as her effort, cf. Bandura, 1977). As mentioned above, a person's outcome control increases with her skill. In this section, we will primarily be interested in the effects that perceived outcome control has on a person's responsibility. The degree to which a person possesses outcome control not only depends on her capacities but also on counterfactual considerations about whether the outcome would have been different had she acted differently (Wells & Gavanski, 1989).

Counterfactual thinking and causal inference Psychologists have found it useful to distinguish counterfactual thoughts in terms of their directionality of contrast. *Upward counterfactuals* are comparisons of the actual world with a somewhat better world and *downward counterfactuals* involve the supposition of a worse world. Several studies have shown that people are more likely to spontaneously engage in upward counterfactual thinking (e.g. Sanna & Turley, 1996). Downward counterfactuals, in contrast, are endorsed comparatively rarely. Accordingly, an outcome's valence – or, more specifically, the affective state motivated by the valence – is one of the main determinants for the activation of the counterfactual thinking process (Roese, 1997). Apart from an outcome's valence, the degree to which the outcome was to be expected has been identified as a promoter for spontaneous causal (Kanazawa, 1992) and counterfactual thoughts (Kahneman & Miller, 1986; Sanna & Turley, 1996). The less an outcome is expected the more likely people engage in causal or counterfactual thinking.

Counterfactuals and responsibility Chapters 2 and 3 have shown that there is a close relationship between counterfactuals, causation and responsibility attribution (Alicke, 2000; Hilton & Slugoski, 1986; Shaver, 1985). In situations in which there

are multiple people involved, a person's control over the outcome is not exclusively determined by their own skill but also by the other people's abilities as well as the way in which the individual contributions are combined to determine the outcome. Gerstenberg and Lagnado (2010) have shown that the same performance can be evaluated differently depending on the group task and the performance of the other players. Their paper provided the first empirical test of a structural model of responsibility attribution developed by Chockler and Halpern (2004). To recap, at the core of this model is a relaxed notion of counterfactual dependence, according to which an event can still be identified as a cause even if changing it would not have made a difference to the outcome in the actual situation. In their model, an individual agent's responsibility for a group's outcome equals $1/(N+1)$, whereby N denotes the minimal number of changes from the actual situation that would have been necessary to generate a situation in which that agent's contribution would have made a difference to the outcome. If no change is needed, the agent receives a responsibility of 1. The more changes would have been necessary to make a person's contribution critical, the more her responsibility decreases.

Consider, for example, our initial darts scenario. In order for the two friends to win the bet, at least one of them needs to hit the centre region. In a situation in which both players hit the centre, their win is overdetermined. That is, the outcome does not depend on either of the players' individual action and hence, a simple *but for* counterfactual analysis would not identify either of them as a cause for the positive outcome (cf. Spellman & Kincannon, 2001). Each player's contribution would only have made a difference to the outcome, if the other player had not hit the centre. Expressed in terms of the structural model of responsibility attribution each person required one change from the actual situation to be pivotal and should hence receive a responsibility of 1/2. Thus, the model predicts that a player's credit should be reduced if the other player hit the centre as well.

5.2.1 Experiment

In order to assess how people attribute blame and credit in a group setting as a function of the players' underlying skill and actual performance, we used the context of a game show environment in which players participated in a team challenge whose outcome affected their individual payoff. The game was similar to an ordinary darts game (see Figure 5.5). It consisted of two phases: first, a practice phase in which each player was given 20 practice shots and, second, the crucial team challenge in which two randomly chosen players formed a team. The team won their challenge if at least one of the two players managed to hit the dart in the centre region. Participants were told that the players differed in terms of how skilled they were in the task. The practice shot patterns were used to manipulate the players' skill levels (see Materials).

The participant's task was to indicate to what extent each player was responsible

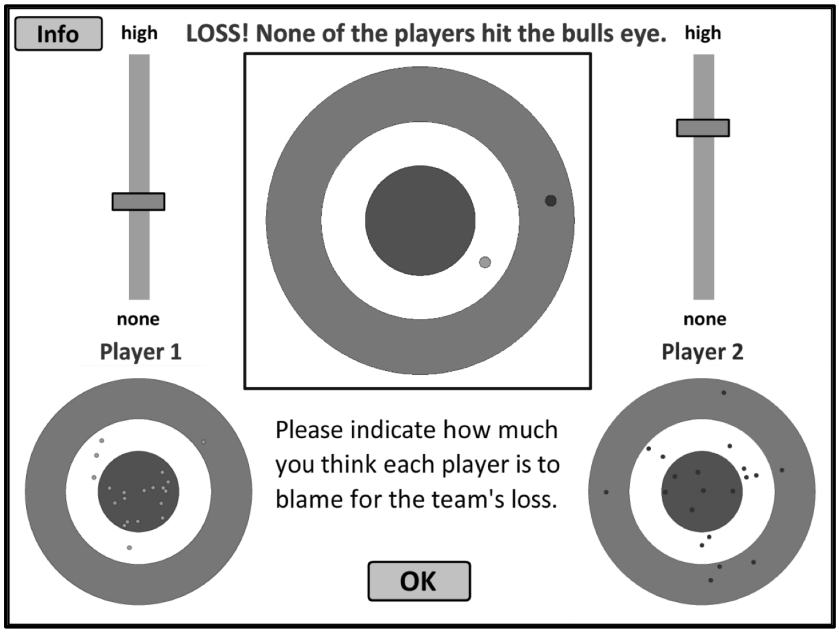


Figure 5.5: Screenshot of the game. Each player's performance in the practice is shown in the bottom corners. The left player's skill level is high and the right player's skill level is low. The performance in the crucial team challenge is shown in the centre.

for the team's result. Participants attributed blame to each player if the team lost and credit if it won. They were informed that their ratings would affect the player's payoff. The more blame a player received for the team's loss, the more his payoff was reduced. The more credit a player received for her team's win, the more her payoff was increased.

We hypothesised that both a player's perceived outcome control as well as the actual performance would influence participants' responsibility attributions. The control factor, as manipulated through the player's skill level, directly influences the availability of counterfactual alternatives (Girotto et al., 1991). If a skilled player missed the centre region, the alternative event in which she did hit the centre is highly available. Likewise, if an unskilled player hits the centre, the alternative event in which he misses is readily available.

Given the prevalence of spontaneous upward over downward counterfactual thinking, we expected the influence of counterfactual alternatives to be stronger in the case of outcomes with negative valence. Accordingly, the skilled player would be blamed more for a loss than the unskilled player. Since downward counterfactual thinking has rarely been shown to occur spontaneously, we expected only a small influence of the skill factor for outcomes with positive valence.

Furthermore, we expected that participants' blame and credit ratings would vary as a function of actual performance. Despite the fact that the rule of the game employs a clear cut-off point in that it only matters whether or not a player hits the centre region, we expected that blame ratings for losses would increase with an increased distance of

a shot from the centre.[5] For wins we expected that players would receive most credit if they hit the centre. Furthermore, we expected that players would receive only minimal credit if they did not hit the centre and that credit ratings would be higher the closer they were to the centre.

Finally, taking the considerations of the structural model of responsibility attribution into account, we expected that a player's responsibility rating would vary as a function of the other player's performance. More specifically, we expected a player's credit rating to be reduced for cases in which the outcome was overdetermined. Hence, a player should receive less credit if the other player also hit the centre as compared to situations in which the other player missed.

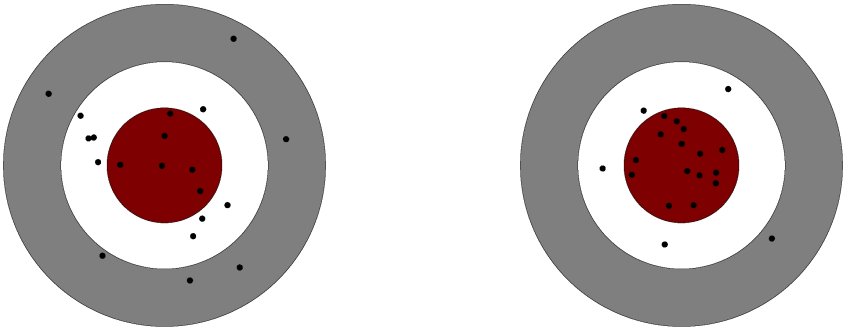
5.2.1.1 Method

Participants 52 participants (31 female) were recruited through the UCL subject pool and took part to receive course credit points or for the chance of winning Amazon vouchers worth £60 in total. The mean age was 23.9 years ($SD = 6.3$).

Design The experiment employed a 3 (skill levels: both players unskilled, both player skilled, one player skilled and one player unskilled) x 9 (performance patterns: full permutation of 3^2 possible shots (centre, medium, outside region) for pairs of players) within-subjects design (see Figure 5.7). Given that the team wins if at least one of the players hits the centre region, this design resulted in 15 cases in which a team won and 12 cases in which they lost.

Materials Figure 5.6 shows an example of the practice shot patterns by the unskilled player and the skilled player. A prototype was generated for each skill level by sampling 20 data points from two centred independent Gaussian normal distributions for the x-axis and y-axis. The skill was manipulated by varying the variance of the distribution. For the unskilled player pattern, 6 shots hit the centre, 8 shots the middle, and 6 shots the outside region. For the skilled player, 15 shots hit the centre, 4 shots the middle and 1 shot the outside region. Hence, based on the practice pattern, the unskilled player had a 30% chance and the skilled player had a 75% chance of hitting the centre. From the prototypical skill patterns, we generated 27 patterns by independently rotating the individual shots around the centre. This procedure ensured that the practice patterns of different players with the same skill level were matched with respect to the most important characteristics. The summed distance of the shots to the centre as well as the number of shots in the different regions was held constant. Nevertheless, the practice patterns still looked different between the players. The patterns of shots for the crucial

[5] This is related to the notion of *changes away from pivotality* discussed in Chapter 4 and the finding that participants' responsibility attributions varied as a function of the distance from the cut-off criterion in the triangle game as well (Gerstenberg & Lagnado, 2010).



(a) unskilled player **(b)** skilled player

Figure 5.6: Prototypical practice shot patterns for the unskilled and skilled player.

team challenge were created in a similar fashion. One prototypical centre, middle and outside ring shot was created and randomly rotated for each pattern in the experiment. Hence, the actual distance of a centre, middle or outside shot was identical in each of the combinations.

Procedure The study was carried out online.[6] At the beginning of the experiment, participants were instructed that they would take the role of a judge in a game show and that their task will be to evaluate different players' performances. The nature of the practice trials and the rules for the team challenge were described as explained above. Participants were told that the players in the game show differed with respect to how skilled they were. As a manipulation check, we showed them 3 patterns of practice shots after the initial instructions and asked them to indicate how skilful each of the players was in the task. Participants used a slider for each skill pattern ranging from -10 ('very unskilled') to $+10$ ('very skilled'). The mean ratings for the unskilled player were $M = -3.1$ ($SD = 3.2$), for the medium skilled player $M = 1.4$ ($SD = 2.5$) and for the skilled player $M = 6.1$ ($SD = 3$). In the main part of the experiment, only the patterns of the unskilled and skilled player were used.

After the skill manipulation check, participants did one practice trial in which the different components of the screen were explained. By clicking on an 'Info' button which remained on the screen throughout the experiment, participants could always remind themselves of the most important aspects of the task. After the practice trial, participants answered a series of 4 forced choice comprehension check questions. On average, participants answered 75% of the questions correctly. After having given an answer, the correct solution was displayed. Participants then proceeded to the main stage of the experiment, in which they evaluated the performance of 27 teams of different players. They always saw each player's performance in the practice trials first and then

[6] A demo of the experiment can be accessed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/skill_demo.html`

the result in the team challenge was revealed. If one of the two players hit the centre region, the team won the challenge, otherwise they lost. Participants were informed about the result of the challenge at the top of the screen. To identify the different players, their shots were coloured differently. If the team won the challenge, participants attributed credit to each player. If the team lost the challenge, participants attributed blame. The sliders ranged from 0 ('none) to 10 ('high').

At the end of the experiment, participants saw the practice patterns and shots in the team challenge for four individual players sequentially. Two players were skilled and two players were unskilled. For each of the skill levels, one of the players hit the centre and one of the players hit the outside ring. For each of the four patterns, participants were asked to indicate how much the following factors influenced the player's result on the final test shot. The factors were: 1) The player's skill level, 2) The player's effort, 3) The pressure of the situation, 4) Chance and 5) The intention to perform this shot. Participants made their ratings on separate sliders ranging from 0 ('not at all') to 10 ('very much'). This final stage was used to gain insight into how participants might explain the different results based on the factors provided. Finally, participants were asked to provide their age and gender.

5.2.1.2 Results

For all statistical tests, we have adopted a significance criterion of $p < .05$ (two-sided). Blame ratings for losses and credit ratings for wins were analysed separately.

Figure 5.7 shows the mean credit and blame attributions for the 27 different patterns used in the main stage of the experiment. The first row shows situations in which both players were unskilled, the second row in which they were both skilled and the third row shows the results for the mixed challenges. We analyse, in turn, the effects of actual performance and underlying level of skill on responsibility attributions.

The influence of performance First, we wanted to see whether blame and credit ratings varied as a function of performance. Indeed, credit ratings were significantly influenced by performance, $F(1, 51) = 428.18, {\eta_p}^2 = .894, p < .05$. Players received most credit if they hit the centre region ($M = 8.77$, $SD = 1.39$). Furthermore, credit ratings for players that did not hit the centre ($M = 2.09$, $SD = 1.81$) were significantly greater than zero, $t(51) = 8.32, p < .05$. Players received significantly more credit if they hit the middle ring ($M = 2.58$, $SD = 2.04$) compared to the outside ring ($M = 1.59$, $SD = 1.75$), $t(51) = 6.07$. Similarly, blame ratings were influenced by the performance of the player as well. A player received more blame if she hit the outside ($M = 6.53$, $SD = 2.23$) compared to the middle ring ($M = 5.55$, $SD = 2.16$), $t(51) = -4.94, p < .05$.

To test how a player was evaluated depending on the performance of the other player, we compared how much credit a player received for a shot in the centre region if the other player also hit the centre or not. A player's credit for a centre shot if the other

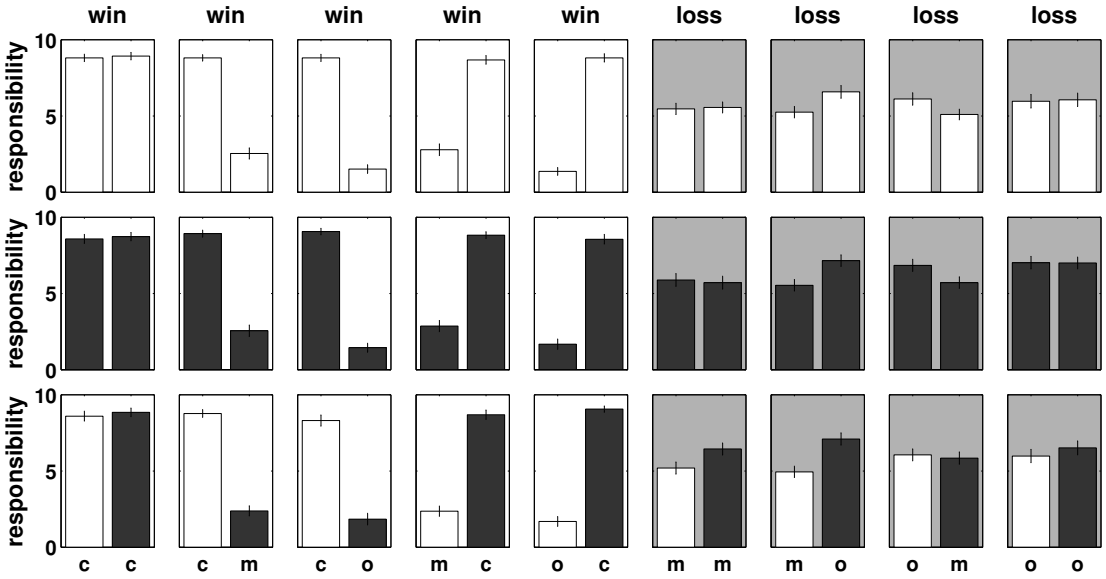


Figure 5.7: Mean credit (white background) and blame (grey background) attributions for the 27 patterns used in the experiment. White bars = unskilled player, black bars = skilled player; c = shot in the centre region, m = middle region, o = outside region. Error bars indicate ±1 *SE*.

player also hit the centre ($M = 8.75$, $SD = 1.67$) was not significantly different from situations in which the other player did not hit the centre ($M = 8.77$, $SD = 1.34$).

The influence of skill Second, we wanted to see whether the blame and credit ratings differed as a function of the player's skill levels. Overall, skilled players received more blame for the team's loss ($M = 6.4$, $SD = 2.23$) than unskilled players ($M = 5.69$, $SD = 2.29$), $t(51) = -2.87, p < .05$. However, there was no significant difference between the credit ratings for skilled players ($M = 6.13$, $SD = 1.07$) and unskilled players ($M = 6.05$, $SD = 0.98$), $t(51) = -0.88, p < .05$.

To look more closely at the effect that the skill level had on people's attributions, we compared the situations in which both players' performance was identical but their skill level differed. Table 5.2 shows that the proportions of participants that either gave equal ratings to both players in these cases or favoured one player over the other differed significantly, $\chi^2(4, N = 52) = 16.83, p < .05$. The majority of participants attributed

Table 5.2: Proportions of participants who either gave identical ratings in mixed-skill challenges with identical performance, favoured the unskilled or skilled player.

	identical	**unskilled**	**skilled**
both centre	34	9	9
both middle	17	9	26
both outside	18	12	22

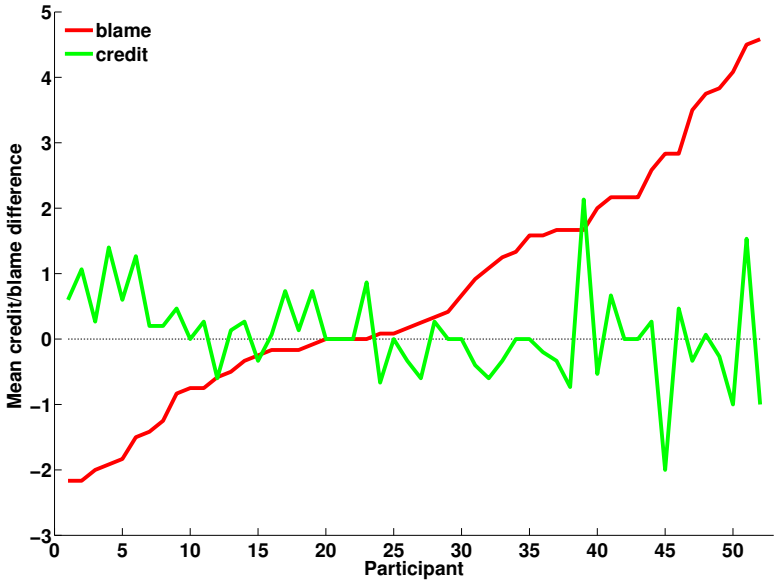


Figure 5.8: Individual differences in the effect of skill on blame/credit attributions. Positive values = skilled player is favoured, negative values = unskilled player is favoured.

credit equally when both players hit the centre. However, in situations in which the team lost and both players either hit the middle or the outside ring, a majority of participants assigned more blame to the skilled compared to the unskilled player.

Figure 5.8 shows the effect of the skill manipulation for each individual participant separately for blame and credit. Positive differences mean that a participant attributed more blame/credit to the higher skilled player. For losses, 29 participants attributed more blame to the skilled player, 19 participants less and 4 participants gave equal blame. For wins, 17 participants attributed more credit to the skilled player, 24 participants less and 11 participants gave equal credit. Participants' blame and credit attributions were negatively correlated as a function of skill, $r = -.34$. Hence, the more blame a participant attributed to a skilled player compared to an unskilled player, the more she credited the unskilled player compared to the skilled player.

Figure 5.9 shows to what extent participants perceived different factors to be important for explaining the players' results for the four test cases at the end of the experiment. We will only discuss the results descriptively. Participants considered the 'Skill' factor to be most important for explaining the shot in the centre by the skilled player. 'Effort', 'Pressure' and 'Intention' varied as a function of performance. For good performances, participants assumed that the player put in high effort, resisted the pressure of the situation and intended to bring about the outcome. The reverse pattern was found for bad performances. The 'Chance' factor varied as a function of expectation. The mean rating for the skilled person hitting the centre was lowest and the rating for the unskilled person hitting the centre highest.

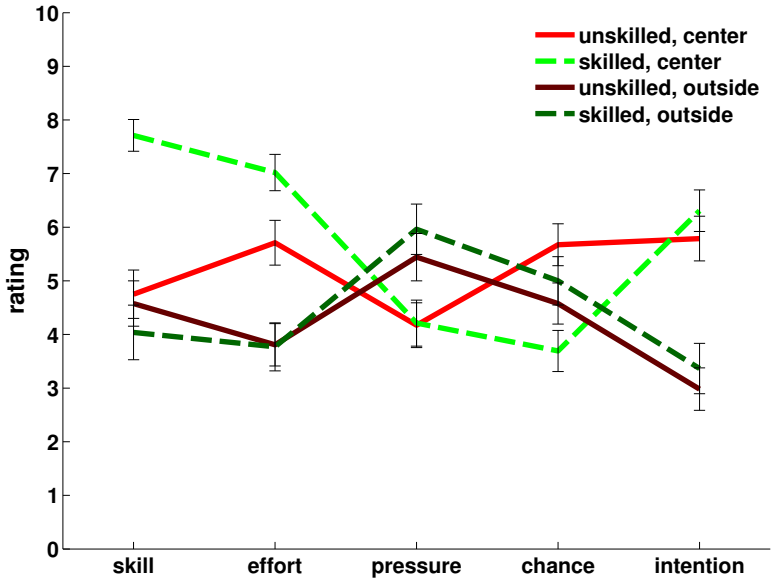


Figure 5.9: Mean ratings indicating how much different factors were seen as having contributed to a shot. Error bars indicate ±1 *SE*.

5.2.2 Discussion

In a novel paradigm in which we systematically varied the skill and performance of agents in a group task, we found that both factors significantly influenced participants' responsibility attributions.

The results revealed that the quality of performance influenced both blame ratings for losses and credit ratings for wins. The worse a person performed the more blame he received for the loss and the less credit for a win. Players received marginal credit for the team's win even if they did not hit the centre region. How a player's performance was evaluated did not vary as a function of the teammate's performance. The influence of skill on responsibility attributions was asymmetric. Skilled players received more blame than unskilled players for losses but credit attributions for wins did not differ significantly as a function of skill.

While the finding that responsibility attributions vary as a function of performance is quite intuitive, the fact that attributions to an individual were not affected by their teammate's performance is surprising. As reported in Chapter 4, we have found in several experiments that participants are sensitive to the performance of the other players and the way in which individual contributions translate into the group's outcome (Gerstenberg & Lagnado, 2010; Lagnado et al., accepted; Zultan et al., 2012). One important difference between the studies reported in Chapter 4 and this experiment concerns the reward function. For example, while in the triangle game the team was rewarded as a whole (Gerstenberg & Lagnado, 2010), participants in the current study were instructed that their blame and credit ratings would affect the payoff of individual players directly. This instruction might have encouraged participants to consider the

players independently and hence no reduction of credit was observed if both players performed well and the outcome was overdetermined.

Another interesting finding concerns the asymmetric effect of the skill manipulation on participants' responsibility ratings. This partly replicates Weiner and Kukla's (1970) finding that reward did not vary as a function of ability (at least for very good outcomes) but punishment did. It is also in line with previous research in the counterfactual literature that showed that counterfactual thoughts are more likely to be spontaneously elicited for outcomes with negative as opposed to positive valence. The fact that the counterfactual alternative in which the skilled player, who exerts more control over the outcome, hit the centre region is more easily available serves as a possible explanation for the increased blame ratings in these situations. If violations of expectation were the main driving force of attributions independent of the valence of the outcome, one would have also expected an increased credit rating for the unskilled player. However, our asymmetric results speak against this explanation.

It is likely that the influence of the skill manipulation would have been even stronger if we had chosen a sample for the patterns of shots in the team challenges that was representative of the players' skill levels. The fact that skill level in the practice and level of performance in the team challenges were independent due to our balanced experimental design might have led some participants to disregard the skill manipulation.

One of the features of our paradigm is that the effect of different combination functions on people's responsibility attributions can be investigated. In our setup, only one of the players needed to perform well in order for the team to win. However, a situation in which both players' good performance is needed is more likely to make participants view the players as a team and hence stronger effects of one player's skill and performance on the other player's evaluation are to be expected.

As discussed at the end of Chapter 4, varying the skill level of individuals within a group has potential implications for how critical each player is perceived for the group outcome and for how close a possible world is to the actual one in which a player of interest would have been pivotal. One of the main potential reasons why there were no strong effects of one player's skill level on another player's responsibility is that our experiment did not explicitly inform participants about the knowledge states of the individual players in a group. As we have seen in Section 5.1, participants' attributions are sensitive to the knowledge states of the players in the team. While we found strong effects of earlier players' performance on responsibility attributions to the last player when that player *knew* about the scores of her teammates, these effects disappeared when participants were instructed that the players do not know about each other's scores. Thus, if many of our participants adopted the reasonable default assumption that the two players don't know how their partner performed in the practice trials when attempting the team challenge, it is not surprising that between-player effects were minimal in this setup. Hence, we should not be discouraged: the issue of how different

priors on variables affect attributions of responsibility remains an intriguing question for future research that can help to refine the core notions of criticality and pivotality.

5.3 Intentions

After having seen empirical demonstrations of how knowledges states (Section 5.1) and performance expectations (Section 5.2) influence attributions of responsibility, we will now look at the role of intentions. As we have seen in Chapter 2, intentions are the distinguishing factor between personal and impersonal causation according to Heider (1958) (see also Malle, 1999) and they feature prominently in theoretical frameworks of responsibility attribution (e.g. Shaver, 1985; Shultz & Schleifer, 1983).

Malle and Knobe (1997) argued that the folk concept of intentionality is closely related to the concepts of skill and awareness (or knowledge) which we have discussed in the two previous sections. Figure 5.10 shows a diagram of how Malle and Knobe model the ordinary people's conception of intentionality. Accordingly, the presence of five components is required in order for people to judge that an action has been performed intentionally. The actor has to have the *desire* to bring about a certain outcome, the *belief* that performing a certain action will result in achieving the desired outcome, the *intention* to perform that action, the *skill* to perform the action (cf. Guglielmo & Malle, 2010; Knobe, 2003b; Nadelhoffer, 2005)[7] and the *awareness* (or knowledge) of performing the intended action.

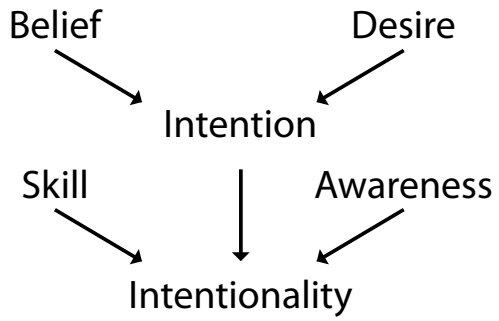


Figure 5.10: A model of the folk concept of intentionality according to Malle and Knobe (1997).

In criminal law, the concept of intentional action is of central importance. The key components that are required for convicting a criminal offender are i) guilty conduct by the defendant (*actus reus*), ii) a guilty state of mind of the defendant (*mens rea*, e.g. the intention to bring about something negative) and iii) the absence of any valid defence (Dickson & Theobald, 2011). Thus, the same outcome (e.g. a dead person)

[7] While Knobe (2003b) has argued that people ascribe intentionality to an action that brings about a morally negative event even when the person lacks skill, Guglielmo and Malle (2010) have shown that when the scenario descriptions are constructed more carefully, differences in skill influence judgements of intentionality even for outcomes with negative moral valence.

brought about by the same *actus reus* is treated differently depending on whether or not a negative intention (*mens rea*) was present. In English law, the classic definition of murder by Coke is: "The unlawful killing of a reasonable creature in being under the Queen's peace with malice aforethought." (cited in Dickson & Theobald, 2011, p. 197) The distinguishing factor between murder and manslaughter concerns the *mens rea* of "malice aforethought". Accordingly, for a conviction of murder it needs to be established that the defendant intended to kill the victim (or intended to cause grievous bodily harm). If the *mens rea* is absent, the defendant can 'only' be convicted of manslaughter.[8] Whereas the degree of punishment for *murder* is necessarily life imprisonment, in the case of manslaughter, the sentencing is at the judge's discretion. Thus, we see that intentions matter a great deal for legal punishment decisions.

However, the law also distinguishes between situations in which the same intention was present (e.g. the intention to kill) but the outcome was different. In United States law, a person can be convicted for *attempted murder* when he or she physically tried to kill someone. A conviction for *murder* requires, of course, that the attempt was successful. Generally, *murder* is punished more severely than *attempted murder*. One obvious reason for why outcomes are of core importance for the law is that they are often objectively assessable: a negative event either happened or not. Intentions, in contrast, are much more difficult to assess. Generally, we have to infer a person's intentions from their actions and people can try to deceive others about what their real intentions were.

In children's development of moral reasoning (Kohlberg, 1983; Piaget, 1932), there is a clear developmental trend from an outcome-based evaluation of behaviour in an early age to a more intention-based evaluation in later years. Hence, a young child might judge a protagonist who accidentally breaks three plates when wanting to help her mom set the table more harshly than a protagonist who intentionally smashes one plate on the floor. In recent years, this developmental trend has been supported through diverse experimental approaches ranging from evaluations of protagonists in scenarios (Nobes, Panagiotaki, & Pawson, 2009) to interactions in simple economic games (Sutter, 2007).

In our everyday life, the valence of intentions and outcomes is correlated: when we intend to do something good we tend to bring about something positive and an intention to do something bad usually results in a negative outcome. However, intentions and outcomes are not perfectly correlated. Sometimes a good intention can bring about a bad outcome (e.g. when a birthday present is not appreciated) and a bad intention can result in a good outcome (e.g. when a malicious plan ends up making the other look very smart and oneself stupid).[9]

[8] The English law distinguishes further between *voluntary manslaughter* for which *mens rea* was actually present but mitigating circumstances reduced full culpability, and *involuntary manslaughter* which includes the unlawful killing of a person without *mens rea*.

[9] A popular German proverb says: "Wer anderen eine Grube gräbt fällt selbst hinein." (The closest proverb in English is: "Harm set, harm get.")

In this section, we will see two studies that employed experimental paradigms in which intentions and outcomes can mismatch. Sometimes, a good intention can lead to a bad outcome and vice versa. The focus will be on the extent to which attributions of responsibility (Section 5.3.1) and monetary interactions (Section 5.3.2) are influenced by an agent's underlying intention versus the resulting outcome.

5.3.1 Intentions, outcomes and responsibility (Gerstenberg et al., 2010)

At the beginning of the movie 'Naked Gun 2 1/2', police officer Frank Drebin is honoured at the presidential dinner for his recent achievement of having eliminated his 1000^{th} drug dealer (see quote at the beginning of this chapter). In response to this, Mr Drebin admits that he had run over the last two men with his car. Luckily, it turned out that they were wanted drug dealers. Cases of 'moral luck' have attracted the attention of philosophers (Nagel, 1979; Williams, 1981), legal scholars (Hart & Honoré, 1959/1985), and psychologists (Mitchell & Kalb, 1981) alike. These situations are characterised by the fact that the outcome of an action influences its moral evaluation retrospectively, even if this outcome was to a large extent beyond the control of the agent. Mr Drebin, for example, receives praise for his reckless driving only because the men he ran over happened to be drug dealers: a circumstance which was clearly beyond his control.

That people are influenced by outcome knowledge is a well established psychological finding (Baron & Hershey, 1988; Fischhoff, 1975). Fischhoff (1975) showed that people are prone to a *hindsight bias*: knowledge about the real outcome influences the perceived likelihood of different possible outcomes. People appear to be unaware of the influence that outcome knowledge exerts on their judgments and are, hence, unable to control for its effect. Baron and Hershey (1988) showed that outcome knowledge influences how people evaluate decisions made under uncertainty. Even when participants had all information relevant to the decision, including the probabilities of each possible outcome, knowledge of the actual outcome nevertheless influenced their judgments of the competence of the decision-maker. Interestingly, when asked whether they *should* take the outcome into account, most participants answered in the negative.

Differential evaluations of identical decisions or actions are also reflected in the Law's differential treatment of negligence versus negligence that leads to harm, as well as cases of attempted murder versus murder. The latter cases share the fact that the person had the intention to kill; however, only in the case of murder did the intended event come about. As seen in Chapter 2, psychologists have shown that intentions play a significant role when it comes to attributions of responsibility (Lagnado & Channon, 2008) and intentionality thus constitutes an important building block of psychological frameworks of responsibility attribution (Alicke, 2000; Shaver, 1985).

The importance of the concept of intentionality has also been recognised by economists. Variations of classic economic games, like the ultimatum game, have been employed to

investigate the effects of outcome versus intention on people's perception of fairness. In the ultimatum game (Güth, Schmittberger, & Schwarze, 1982), a first player is allocated a certain amount of money. She can then decide how much of that money to give to a second player, who can either accept or reject the offer. If he refuses, both players get nothing. Two main findings with respect to the influence of intentions on the behaviour of the second player are worth mentioning. First, if the allocation of the first player is determined by a computer and hence cannot be ascribed an intention, the rejection rates for 'unfair offers' are significantly lower (Falk, Fehr, & Fischbacher, 2008). Second, the rejection rates of unfair offers strongly depend on the allocator's set of possible alternatives (McCabe, Rigdon, & Smith, 2003). The acceptability of an action is hence evaluated with respect to the choice set and an unequal offer more readily accepted if the allocator could not have been kinder. In order to accommodate these findings, economists have moved from theories of fairness that only consider distribution of outcomes (Fehr & Schmidt, 1999) to theories based on intentions (Dufwenberg & Kirchsteiger, 2004) and theories incorporating both intentions and outcomes (Falk & Fischbacher, 2006).

As demonstrated by the moral luck example in 'Naked Gun 1/2', there is another factor beyond intentions and outcomes that is relevant when it comes to considerations about fairness or attributions of responsibility: the control an agent has over the outcomes he intends to brings about. Our environment is fundamentally noisy and, most of the time, we only have partial control over the effects of our actions. While it is true that the valence of intention and outcome are correlated in everyday life, this relationship is imperfect. Good intentions can sometimes lead to bad outcomes and bad intentions to good ones. In order to understand the complex relationship between intentions, outcomes and control it is necessary to create experimental situations in which these factors can be dissociated.

In a recent study, Cushman, Dreber, Wang, and Costa (2009) investigated the effects of intention versus outcome on perceived fairness in a two-player, allocater-responder game. Similar to the ultimatum game, the allocator proposed how a pot of $10 should be shared. Allocations were either stingy (player 1: $10; player 2: $0), fair ($5; $5) or generous ($0; $10). The responder could punish or reward the allocation of player 1 by subtracting or adding up to $9 to her account. Importantly, in one condition of the experiment, the allocator only had partial control over the outcome. She had to choose which one of three possible dice she wanted to roll. These dice differed in terms of the probability with which they led to stingy, fair or generous outcomes. Following a strategy format (Selten, 1967), responders had to indicate for each of the 9 possible combinations (e.g. generous die, stingy outcome) how much money they wanted to add or subtract from the allocator. The results revealed that participants were much more influenced by actual outcomes than by intentions. Responders tended to subtract money for selfish outcomes for all three dice, whereas they added money for fair and

generous outcomes. The choice of die exerted only a small effect on this general pattern. Surprisingly, the results of a condition in which the allocator had perfect control were virtually identical. Hence, the study provides further support for the finding that people can be so sensitive to outcomes that they sometimes disregard the underlying intention that lead to that outcome. However, Cushman et al. (2009) admit that methodological limitations might have contributed to their findings. Importantly, since the responder is part of the game, it is the outcome and not the intention that is the most relevant to him. In order to validate their findings, it is important to investigate how an independent judge would have decided.

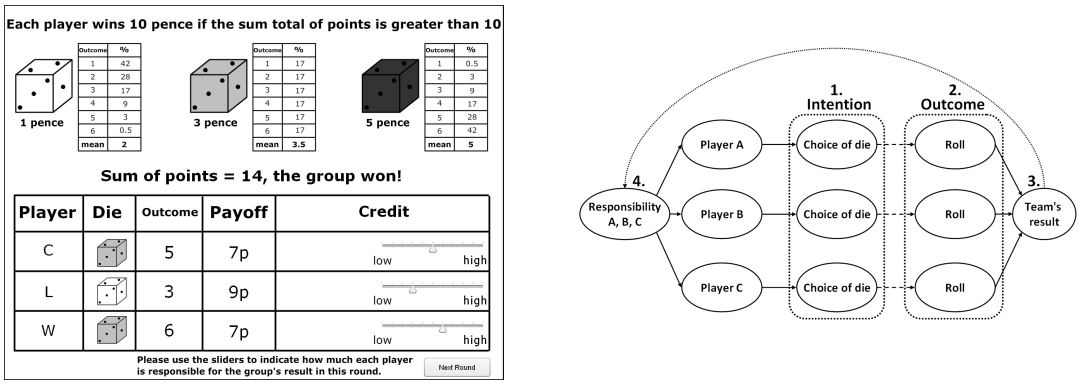
The current experiment addressed this limitation. We created a setting in which an external observer evaluated the behaviour of agents participating in an experimental game. The following scenario helps to exemplify the main components of our experiment. Sarah is running for the position of student representative. Three friends are helping her campaign by distributing flyers. Tom puts in a lot of effort and distributes 100 flyers. John puts slightly less effort into the campaign and only distributes 50. Finally, Alex thinks that Tom's and John's contributions are probably already enough to win the campaign and he only distributes 30 flyers. As it turns out, 20 of Tom's, 20 of John's and 25 of the people who received their flyer from Alex voted for Sarah. As a result, Sarah won the election. Assuming that Sarah knows about both the number of distributed flyers and the votes she received, how much is she going to praise each of her three friends for their contribution to her win?

Two aspects of the outlined scenario are important with respect to the current study. First, it shows how intention and outcome can sometimes mismatch in situations in which agents exercise only partial control. Despite Tom's good intention and effort he contributed no more to the collective outcome than John and even less than Alex. Second, the scenario entails a component that is characteristic of social dilemmas (see, e.g. Hardin, 1968). Each individual agent has to weigh the cost of the effortful process of distributing flyers with the potential gain of an election won. Alex's thought process indicates each person's motivation to free-ride on the effort of the others. Assuming the spoils of a victory are equally shared, the person who put in the least effort will have the highest net benefit.

The current study investigated the effects of intended and actual contributions on responsibility attribution in a group context in which agents had only partial control over their contributions. How well can intended contributions, actual contributions, or their combination explain participants' responsibility attributions?

5.3.1.1 Experiment

The aim of the experiment was to generate a situation in which intended versus actual contributions could dissociate. Participants took the perspective of an external observer



(a) Screenshot of the game **(b)** Structure of the game

Figure 5.11: (a) Screenshot of the game and (b) underlying structure. Numbers 1. – 4. indicate the different components of each round.

and judged the behaviour of computer players engaging in an experimental game (see Figure 5.11a).

In each round of the game, three computer players were randomly selected to form a group. Each player chose one die out of the three available types of dice to roll. Several players could choose the same type of die if they wanted to. The dice differed in terms of their underlying probability distributions (see top part of Figure 5.11a). The white die had a higher probability of smaller values, the grey die was fair, and the black die was skewed towards higher values (in the experiment, the colours were bronze, silver and gold). The group of players won a round if the sum of their outcomes was greater than 10. If the group won a round, 30 pence were equally distributed between the players. If the group lost, no money was distributed. Importantly, the players had to pay different amounts for the dice before they rolled them. The white die cost 1 pence, the grey die 3 pence, and the black die 5 pence. Each individual player's payoff was a function of the group's result, that is, whether they won or lost, and the money he had to pay for the die of his choice. For example, in the situation depicted in Figure 5.11a, player C chose the grey die and threw a 5. The cost of the grey die is 3 pence. The team won this round as the summed outcomes of their dice was greater than 10. Hence, each player earned 10 pence for the win and C's payoff in this round was 7 pence (10 pence for the win minus 3 pence that he paid for throwing the grey die). Participants' task as independent judges was to indicate how much they thought each player was responsible for the group's result in each round.

Participants were left uninformed about the fact that the computer players chose each of the dice with an equal probability. The chosen payoff function created a social dilemma. The overall probability of winning was 50%. The probabilities of winning given that a player had chosen the white, grey or black die were 33%, 50% and 68%, respectively. This led to an expected payoff of 2.3 pence per round for the white die

($33\% \times 9 - 67\% \times 1 = 2.3$). The expected payoffs for the grey and black die were 2 and 1.8 pence. Hence, there was an incentive for each player individually to choose the white die. However, if all of the players chose that die, the probability of the team winning was only 2%, and the expected payoff -0.8 pence.

Figure 5.11b shows the underlying structure of the experiment. The choice of die reflected the *intended* contribution of the player while the team's result was a function of the *actual* contributions. We predicted a main effect of intention: the same outcome of roll will elicit different responsibility attributions dependent on the choice of die. We also predicted a main effect of outcome: responsibility attributions for a given die will vary with the outcome of rolling this die. Finally, based on previous research (Cushman et al., 2009) we predicted that outcomes will affect participants' responsibility ratings more strongly than intentions.

Method

Participants and materials 87 participants from the UCL subject pool participated for the chance of winning one of six amazon vouchers worth £150 in total. 61 participants were female and the mean age was 23.1 years ($SD = 5.74$). With the second part of the experiment added at a later stage (see Procedure), 40 participants performed only the first part of the experiment, whereas the remaining 47 participants performed both. The study was conducted online and programmed with Adobe Flash.[10]

Procedure Participants were informed that the experiment would take 20 minutes and that their task was to evaluate the behaviour of players engaged in an experimental game by attributing credit for wins and blame for losses. Participants read a description of the three dice which made it clear that they differed in terms of both probability distribution and price. A practice round served to familiarise participants with the structure of the game. After the practice round, participants had to answer a set of comprehension questions to ensure that they had understood the rules of the game correctly. The game was then played for 20 rounds.

On each round, participants saw a table that showed for each player which die she had chosen, the outcome of having rolled that die and the amount won or lost in that round. Players were indicated by capital letters which changed in each round. This was done to prevent participants from forming an overall impression about individual players. The header above the table showed the sum of points and changed its colour from green to red according to whether the round was won or lost. For each player, participants attributed blame for losses or credit for wins, by moving a slider ranging over a scale from 0–10. Its endpoints were labelled 'low' and 'high'. In line with the

[10] A demo of the experiment can be accessed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/intention_demo.html`

result of the round (loss/win), the label (blame/credit), colour (red/green) and position of sliders (middle to left/ middle to right) of the last column changed.

47 of the 87 participants also completed a second stage of the experiment. Those participants were informed after the twentieth round that they would see 14 novel situations that could have occurred in the game and which were of special interest to the researchers. As explained below, the test cases were chosen so as to enable a fine assessment of the weight assigned to intentions and outcomes. The order of these test cases was randomised. Finally, participants were asked to indicate in a text box whether they had focused on the choice of die, the outcome, or both.

Results

Mean responsibility ratings Figure 5.12 shows the mean responsibility attributions in the first stage of the experiment as a function of choice of die and outcome of roll separately for negative (top) and positive (bottom) group outcomes. We ran a linear mixed effects model which included a random intercept for each participant, and fixed effects for choice of die (white, grey, die), outcome of roll (1, 2, ..., 6) and group result (win, loss). In order to facilitate the interpretation of the results, we reverse-coded responsibility attributions for negative outcomes by subtracting participants' attributions from 10 (the maximum of the scale). There was a significant effect of choice of

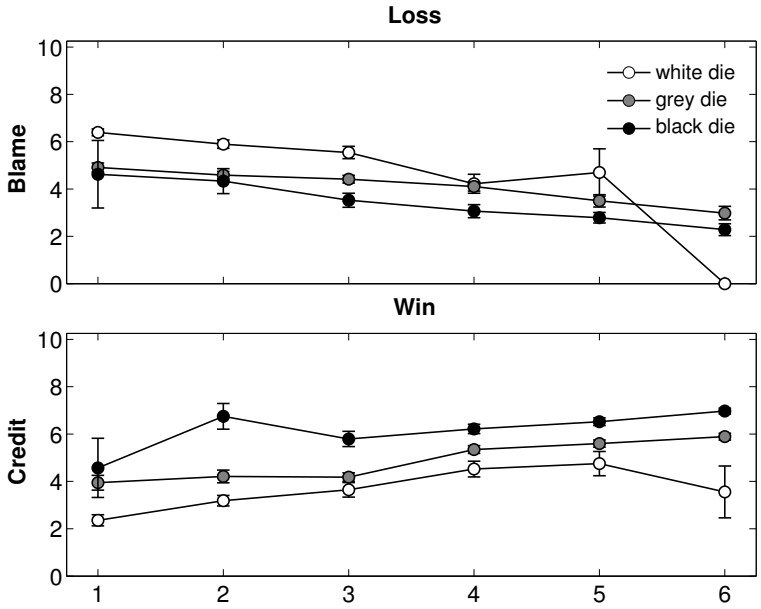


Figure 5.12: Mean responsibility ratings for each combination of die and outcome for losses (top) and wins (bottom). Lines represent the different dice and values on the x-axes indicate the outcome of rolling each die. Error bars indicate $\pm 1 SEM$. *Note*: The data point for the blame attribution for the white die which rolled a 6 is based on a single observation only.

die $F(2, 5161.67) = 23.50, p < .001$, outcome of roll $F(1, 5148.53) = 97.11, p < .001$ and group result $F(1, 5160.66) = 9.14, p = .003$. There were no significant interactions.

The cheaper a player's die the more he was blamed for negative outcomes (white: $M = 6$, $SD = 3.45$; grey: $M = 4.35$, $SD = 2.4$; black: $M = 2.98$, $SD = 2.92$) and the less credited for positive outcomes (white: $M = 3.36$, $SD = 3.03$; grey: $M = 5.21$, $SD = 2.35$; black: $M = 6.66$, $SD = 2.99$). Furthermore, blame attributions decreased and credit attributions increased the higher the outcome of the die roll. Finally, participants' credit attributions ($M = 5.51$, $SD = 3.06$) were higher on average than their blame attributions ($M = 4.84$, $SD = 3.25$). These analyses show that overall, both the choice of die and the outcome influenced participants' responsibility ratings. However, the results cannot reveal how individual participants weighted these two factors. To find out, we conducted individual regression analyses, and report them below.

Regression analysis First, we ran the following three separate regression analyses based on the overall data (87 participants x 20 rounds x 3 ratings data points):

$$\text{intention-based model: } responsibility = \beta_0 + \beta_{1\ die} \tag{5.1}$$

$$\text{outcome-based model: } responsibility = \beta_0 + \beta_{1\ roll} \tag{5.2}$$

$$\text{mixture model: } responsibility = \beta_0 + \beta_{1\ die} + \beta_{2\ roll} \tag{5.3}$$

All three regression models accounted for a significant amount of the variance in the data (see Table 5.3). Overall, the intention-based model accounted for more variance than the outcome-based model. Moreover, when both choice of die and outcome of roll were used as predictors, die choice was a stronger predictor of participants' responsibility attributions than outcome of roll.

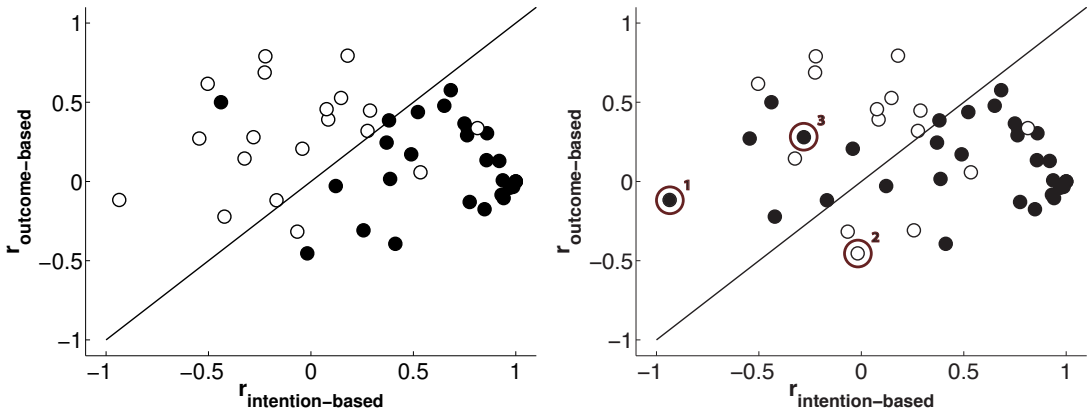
Evaluation of Test Cases To break the results down even further, we ran the regression models for each individual participant. Based on the magnitude of the correlation with the intention-based regression model versus the outcome-based regression

Table 5.3: Results of overall regression analyses.

Model	R^2	F	$p <$	β	t	$p <$
intention-based	.146	892.94	.001	$.382^a$	29.88	.001
outcome-based	.127	758.50	.001	$.356^b$	27.54	.001
mixture	.164	512.81	.001	$.261^a$	15.28	.001
				$.182^b$	10.65	.001

$a = \beta_{die}$, $b = \beta_{roll}$

model, we grouped the 47 participants who completed the second stage of the experiment in two groups. We used two different categorisation procedures based on either (a) the maximum correlation $max(r)$ or (b) the maximum variance accounted for $max(r^2)$. These classification procedures only differ with respect to participants whose attributions are negatively correlated with the choice of die or the outcome of the roll. That is, whereas a participant whose attributions are negatively correlated with the choice of die ($r_{intention\text{-}based} < 0$) and uncorrelated with the outcome of the die roll ($r_{outcome\text{-}based} \approx 0$) would be classified as outcome-based according to the $max(r)$ procedure, she would be classified as intention-based according to the $max(r^2)$ procedure. We used the categorisation to look at how participants attributed responsibility for the chosen test cases (described below). Figure 5.13 shows how well the behaviour of the classified participants was accounted for in the test cases separated for the two different classification procedures.



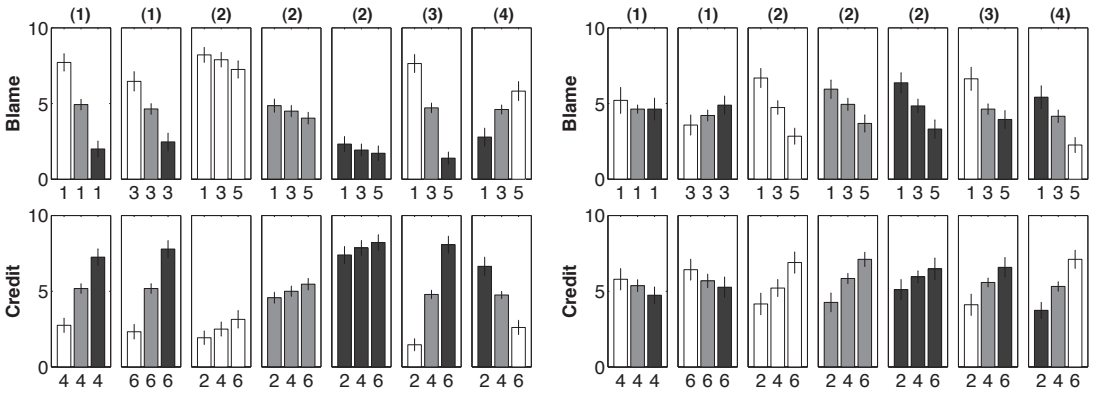
(a) $max(r)$ categorisation **(b)** $max(r^2)$ categorisation

Figure 5.13: Scatterplots of correlations with outcome-based model and intention-based model according to the (a) $max(r)$ categorisation procedure ($N_{intention\text{-}based} = 28$, $N_{outcome\text{-}based} = 19$) and the (b) $max(r^2)$ categorisation procedure ($N_{intention\text{-}based} = 32$, $N_{outcome\text{-}based} = 15$). Black circles indicate participants who were classified as having focused on intentions based on the first part of the experiment. White circles indicate participants who were classified as outcome based. *Note*: The marked cases in (b) are discussed in the General Discussion.

The test cases were constructed to enable a fine analysis of the relative weights assigned by participants to intentions versus outcomes. It should be noted that in the first 20 rounds of the experiment the choice of die and outcome of roll were positively correlated due to the chosen probability distributions ($r = .68$, $p < .001$). In contrast, for the test cases, the choice of die and outcome of roll were uncorrelated ($r = 0$). This shows that these test cases indeed created situations that could be used to distinguish between intention-based and outcome-based participants.

Figure 5.14 shows how participants who were classified as intention-based or outcome-

based according to the $max(r)$ procedure attributed responsibility in the test cases.[11] The test cases can be categorised into 4 groups: (1) different dice, same roll; (2) same dice, different rolls; (3) congruent; (4) incongruent. With 'congruent' we mean that the quality of die and outcome of roll corresponded (i.e. the expensive die led to a high and the cheap die to a low outcome); 'incongruent' means that the quality of die and the outcome of the roll mismatched.



(a) Intention-based participants ($N = 28$) **(b)** Outcome-based participants ($N = 19$)

Figure 5.14: Mean responsibility attributions of (a) intention-based and (b) outcome-based participants for 14 test cases. The categorisation is based on the $max(r)$ procedure. The top row shows blame attributions for losses and the bottom row credit attributions for wins. The values on the x-axes indicate the outcome. The colours of the bars indicate the dice. Error bars indicate ± 1 *SEM*.

Inspection of the graphs validates the original partition. First, in the congruent test cases (3) which serve as a manipulation check, both groups show the same pattern of attributions, with that for the intention group being steeper than that for the outcome group. For the intention test cases (1), the differences in attributions are large for participants in the intention group and small for participants in the outcome group. An opposite pattern of attributions is evident with the outcome test cases (2): there the intention group exhibits small differences and the outcome group exhibits large differences. Finally, and most interesting, the pattern of attributions reverses in the incongruent cases (4). Despite the fact that in these situations the expensive die led to the lowest outcome, the intention-based participants attribute the least blame to this player for the loss (Figure 5.14a, top) and the most credit for the win (bottom). In contrast, the attributions of the outcome-based participants for these cases closely follows the number rolled, independent of the choice of die (Figure 5.14b).

Discussion In this section, we investigated the influence of intended versus actual contributions on the attribution of responsibility in a group context. We found that

[11] The two categorisation procedures agree for 39 out of 47 participants. The choice of procedure does not affect the qualitative patterns of the results.

both intention and outcome exerted a significant influence on participants' attributions. Furthermore, we provided evidence that individuals differ in the extent to which they base their attributions on intentions or outcomes.

Our experimental procedure allowed us to dissociate intentions from outcomes. We found that the majority of participants were better explained as having focused on intended rather than actual contributions. Out of the 49 participants who were better explained by the intention-based model, 24 had a correlation of $r = .8$ or higher. A common strategy for participants focussing on the choice of die was to attribute 0, 5, 10 credit for wins or 10, 5, 0 blame for losses for the white, grey and black die, respectively. Eight participants consistently adopted this strategy. Methodologically, the current experiment shows that it is important to analyse the data on the level of individual participants. While on an aggregate level, it appears that participants are weighting both choice of die and outcome of roll to determine their responsibility attribution (see Figure 5.12), more careful analyses reveal that a large proportion of participants actually tend to either focus on the intention or the outcome (see Figure 5.13).

While the different categorisation procedures that we have used agree for most of the participants, there are some participants for whom the categorisation procedures disagree. Consider the three marked participants in Figure 5.13b. Participant 1's attributions are strongly negatively correlated with the choice of die. That is, this participant blamed players who chose the more expensive die *more* when the team lost and credited them *less* when the team won. In contrast, participant 2's ratings were uncorrelated with the choice of die but negatively correlated with the outcome of the die roll. Hence, this participant blamed players who ended up rolling a *high number* more for the loss and credited them less for the win. Both of these attribution strategies strike us as somewhat odd and it is worth noting that they were hardly adopted by anyone else. While for these more extreme cases, the $max(r^2)$ classification scheme might be more intuitive (participant 1 clearly cared about the choice of die and not the outcome of the roll), this procedure is problematic for participants whose attributions were more sensitive to both factors. For example, it is questionable whether a participant whose attributions have a slightly higher negative correlation with the choice of die than they have a positive correlation with the outcome of the die roll should be classified as intention-based (see participant 3). The problem of categorisation notwithstanding, our results clearly show that participants adopt a variety of different strategies.

At this point, we can only speculate about the underlying factors driving these interindividual differences. Different interpretations of the notion of responsibility could have influenced participants' behaviour (cf. Hart, 2008). Outcome-based participants might have endorsed a *causal* conception of responsibility. On this view, players that rolled high numbers were assigned greater credit because their contributions actually caused the win. Intention-based participants, on the other hand, might have used a *moral* (or control-based) conception of responsibility (cf. Alicke, 2000). Hence, players

were judged for their choice of die which reflected their underlying attitude towards the team. Alternatively, the results could reflect interindividual differences in the ability or motivation to mentalise. We would assume that people who find it hard to take another person's perspective are more likely to focus on the actual outcome rather than the underlying intention. We are planning to use a simplified version of the employed paradigm to test this hypothesis on a patient group with deficits in mentalising. Finally, outcome-based participants' ratings might have been influenced by beliefs about the gambling-competence of players. On this view, rolling a high number with the cheap die reflects a special ability deserving credit. Some of the participants' written comments confirm the influence of such arguably non-normative considerations.

Why did we find a relatively stronger effect of intentions when previous studies postulated the existence of an outcome bias (e.g. Cushman et al., 2009)? Several differences between studies that draw their conclusions from economic games, such as the ultimatum game, and our study could potentially explain these divergent results. First of all, most of the studies in the economic literature were interested in exploring perceived fairness and not directly in responsibility attributions. Although we presume that these notions are tightly linked, it might be that considerations about fairness and responsibility can lead to different results. Second, the participants in those studies directly experienced the outcomes, while inferring the intentions of the other player was not incentivised. This might explain why they found a relatively stronger effect of outcomes compared to intentions. In our study, in contrast, participants acted as independent external judges. It is, hence, less likely that their attention was biased towards outcomes. We will explore some of these differences between Cushman et al.'s (2009) study and our experiment in more detail in the next section which employs an experimental paradigm that is more similar to the one used by Cushman et al.

A feature of the employed experimental paradigm is its potential to explore different combination functions between the individuals in the group. Gerstenberg and Lagnado (2010) have shown that the way in which individual contributions are translated into group outcomes significantly influences people's responsibility attributions. Accordingly, an identical individual contribution can lead to very different responsibility attributions as a function of the group context. While the current experiment used an additive combination function for the contributing players, we will investigate in future experiments how attributions change when the rule of the game reflects a minimum function (i.e., the group wins if no player rolls a 1) or a maximum function (i.e., the group wins if at least one player rolls a 6). It will be interesting to see whether participants are more likely to focus on the actual rather than the intended contribution when the combination function is non-compensatory.

Finally, the current paradigm can be used to explore how uncertainty affects responsibility attributions. In the current version, all information was revealed to the participant. In everyday life, however, we normally do not have direct access to other

people's intentions. Rather, we try to infer others' intentions from their behaviour. Our paradigm allows us to model this situation. Instead of revealing all the information to the participant, we will only show the outcome of rolling the die but not which die the players have chosen. We will explore this effect of uncertainty in a different setup below.

5.3.2 Beyond outcomes (Schächtele et al., 2011)

The previous experiment established that the degree to which people hold a group member responsible for the group outcome depends both on the group members intended outcome and what he or she actually brought about.[12] While most participants took both factors somewhat into account, there was also a significant group of participants that exclusively focused on the underlying intention. We discussed some of the methodological differences that might have contributed to a stronger overall reliance on intentions in our study as opposed to a stronger reliance on outcomes in Cushman et al.'s (2009) study.

In this section, we will use an experimental paradigm which matches Cushman et al.'s (2009) methodology more closely. Rather than asking uninvolved participants for judgments of responsibility, we have participants interact with each other. As mentioned above, the greater reliance on intentions might have been due to the fact that participants did not experience the outcome themselves. Generally, it's easier said than done: telling someone else not to be angry because of a negative outcome which was due to some misunderstanding is easier than when oneself has experienced the negative outcome. In the following experiments, we ask participants to put their money where their mouth is and replace attributions of responsibility with monetary interactions that are consequential for participants. As discussed in Chapter 2 the notion of responsibility is notoriously polysemous and interindividual differences might thus have resulted from differences in the semantic interpretation of the term. Money, in contrast, is unambiguous. Will participants still care about the underlying intentions of an interaction partner or just punish or reward the other person based on the outcome they have received?

In economic theory, the utility functions constructed to describe revealed preferences (Samuelson, 1938) can incorporate anything, including perceived intentions. In practice, most models of utility are just based on personal income. However, more recently, due to an increasing interaction between psychology and economics (Rabin, 1998), researchers have begun to develop formal models of social preferences which are not solely based on individual outcomes but include considerations of fairness (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999). Although these theories are different from pure income-maximisation models, they remain concerned with outcomes, namely distributions of payoffs. A third generation of models goes further still and adds perceived intentions to

[12]This section has greatly benefited from Simeon Schächtele's contribution who was the first author of the corresponding article. Parts of this section are reprinted from Schächtele et al. (2011) and have been written in collaboration with Simeon Schächtele and David Lagnado.

the utility function (Charness & Rabin, 2002; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006).

Previous studies on the role of intentions in economic games have employed different experimental designs and have reached different conclusions. A first group of studies compares responses to intentional actions against the responses to random events, whereby the intentional choices and the random events have identical monetary consequences. Falk et al. (2008) find that rejection rates in an ultimatum game (Güth et al., 1982) are significantly lower when the offers are determined randomly by a computer, as opposed to set intentionally by a person. In contrast, Stanca (2010) finds that intentions are irrelevant in a within-subject comparison between random and intentional first-moves of a gift-exchange game. A second group of studies (e.g. Falk, Fehr, & Fischbacher, 2003; McCabe et al., 2003) finds an effect of intentions by comparing responses to identical actions that have been selected from different sets of alternative actions. In an ultimatum game, for instance, a disadvantageous offer is more likely to be rejected when the alternative is an equal split, as opposed to an even less advantageous offer. A third group of studies (Charness & Levine, 2007; Cushman et al., 2009) compares responses to identical outcomes that were reached by different combinations of intentional choices and chance.

Both Cushman et al. (2009) and Charness and Levine (2007) find an effect of intentions, but the relative importance of intentions versus outcomes differs widely between their studies. Cushman et al. (2009, p. 1), conclude that "accidental outcomes guide punishment" whereas Charness and Levine (2007, p. 1061), report that "intention appears to be a stronger force than distribution". Our basic research design follows the general approach of Charness and Levine and Cushman et al. By combining what we see as the desirable features of both studies, we gauge the robustness of their results. Like Cushman et al., our design employs three levels of intentions and outcomes, allowing us to compare negative and positive reciprocity. Like Charness and Levine, however, we abstain from using the strategy method (Selten, 1967). In the strategy method, participants are asked to indicate their responses for all possible situations in a large table all at the same time. Usually, one of the possible situations is then chosen randomly and the participants receive their pay-off depending on the answer that they have indicated for this particular situation. However, the strategy method encourages a rather cold mode of reasoning and facilitates participants to give answers in a consistent fashion. In our experiments, in contrast, we employ a more realistic setup in which participants interact with each other sequentially and experience the outcomes of each interaction directly. Furthermore, we add realism by making responses costly. Thus, just like in real life, punishing or rewarding another person for their actions doesn't come for free. It minimally involves spending time and effort.

In practice, the role of intentions in social interactions is affected by the fact that we lack direct access to other people's intentions. In particular, statements of intention

may be deceptive. Consider an example from the legal domain. A murderer may have a strong incentive to lie to the jury about his intention to kill the victim. As a result, the jury has good reason not to take the stated intention at face value (see Fenton et al., 2012; Lagnado, 2011b; Lagnado, Fenton, & Neil, 2012, for related issues that arise in the context of alibi and witness evidence). In business, negotiation and other social interactions, deceptive statements about intentions exist as well. Research has shown that people are sensitive to the incentives for and the presence of deception (e.g Shalvi, Dana, Handgraaf, & De Dreu, 2011). We incorporate this feature in Experiment 2 and assess how it influences the effect of intentions in participants' responses. Could it be that intentions matter more when they are not subject to the possibility of deception? Notice that from the viewpoint of outcome-based utility, the absence or presence of deception should not matter because intentions are irrelevant.

5.3.2.1 Experiment 1

The central building block of our experiments is a sequential, probabilistic allocator-responder game with three choice alternatives and three outcomes (cf. Figure 5.15, left panel). First, the 'selector' chooses one of three random devices dubbed 'wheels of fortune'. The computer then determines the outcome of the wheel of fortune. Outcomes refer to divisions of 30 tokens ($\approx$ £1) between the selector and the responder. The responder is informed about the selector's choice and the wheel's outcome. Finally, the responder can subtract or add between -15 and $+15$ from the selector's payoff, whereby for 3 tokens added or subtracted she has to give up one of her own tokens.[13] The core idea of this design is that the selector's wheel choices signal her intention towards the responder, who can reciprocate intentions and/or adjust outcomes by adding or subtracting tokens. The probabilities of the wheels and the monetary divisions associated with the outcomes are summarised in Table 5.4.

Table 5.4: Wheels of fortune: Probabilities and outcomes.

Probability	**Wheel 1**	**Wheel 2**	**Wheel 3**	**Payoff**
Outcome 1	60%	20%	10%	20 | 10
Outcome 2	30%	60%	30%	15 | 15
Outcome 3	10%	20%	60%	10 | 20

Note: Payoffs (Selector | Responder) in experimental tokens.

Each wheel choice can result in any of the three outcomes, but the wheels differ in the probabilities of yielding a particular outcome. Wheel 1 has a high probability

[13]The cost ratio of 1:3 for adding or subtracting tokens coincides with the ratio used in experiments on altruistic punishment and is comparable in magnitude to the 1:4 ratio employed in Charness and Levine (2007).

of allocating 20 tokens to the selector and 10 tokens to the responder. Wheel 2 has a high probability of an equal split of the 30 tokens. Wheel 3 has a high probability of allocating 10 tokens to the selector and 20 to the responder. We follow Cushman et al. in labelling both the outcomes and intentions 'stingy', 'fair' or 'generous'. In the experimental instructions, however, we avoided such terms and used neutral language. An example round could look like the following: the selector decides to choose Wheel 2 (i.e. the 'fair' wheel). The wheel comes up on Outcome 1 (i.e. the 'selfish' outcome). If the responder did nothing, the selector would receive 20 tokens and the responder 10. However, let's consider that the responder was unhappy with this outcome and invested one of his tokens to subtract three tokens from the selector. Thus, the selector received

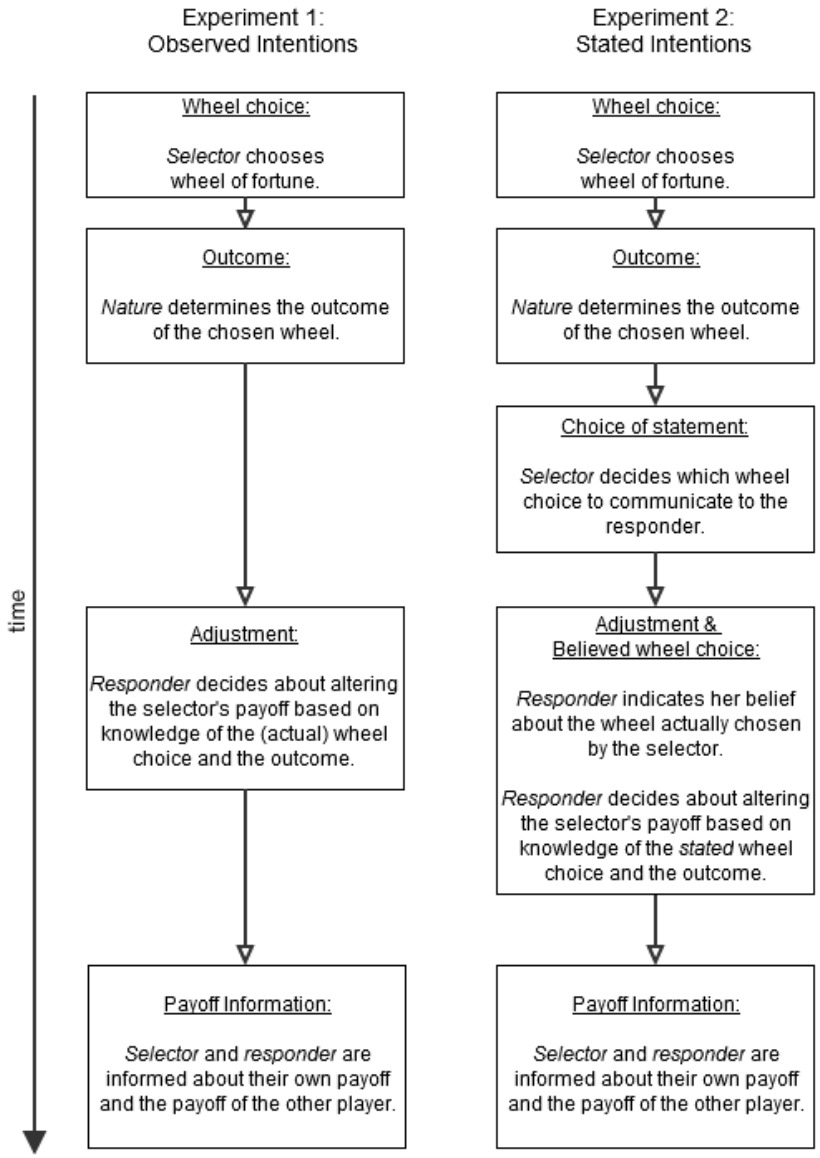


Figure 5.15: Sequence of events in both Experiments.

17 tokens in this round and the responder 9.

The experiment consisted of two parts. In Part A, the wheel of fortune game was played 16 times under a protocol of random matching with anonymity. The roles of selectors and responders were assigned randomly and remained fixed throughout Part A. In Part B, all 16 participants of a session acted as responders. Participants were asked to indicate their responses to all nine possible combinations of intentions and outcomes, which were presented in random order. This resembles the strategy method in that responses are collected for all possible first moves, while the sequential presentation reduces cognitive demands. Participants were informed that their responses in Part B had no effect on actual payoffs. Part B allowed us to collect an equal number of observations for all nine combinations of intentions and outcomes. We included it as an additional check on the results from Part A, in which we had no control over the number of observations per cell.

Method

Participants and materials The experimental session was conducted at the Centre for Economic Learning and Social Evolution (ELSE) of University College London. A total of 16 participants were recruited online from the ELSE subject pool. 11 participants were female and the median age was 22.5 years ($SD = 2.88$). The experiment was programmed with the software package z-Tree (Fischbacher, 2007).

Procedure Instructions for Part A were read aloud by the experimenter and questions related to the instructions were answered in private. After Part A, instructions for Part B were read out aloud and questions answered in private. Before each of the two parts, a comprehension check had to be passed. At the end of the experiment, the participants' earnings were paid out to them in private. Average earnings were £10.91 ($SD = 0.84$), including a show-up fee of £5. The experiment lasted 60 minutes.

Results

Selectors' wheel choices and ex-post profitability The fair wheel was the most common choice ($n = 60$), followed by the stingy ($n = 53$) and the generous wheel ($n = 15$). Due to responders' adjustments, selectors choosing the stingy wheel earned less on average than the fair wheel (10.11 vs. 12.35 tokens), which in turn earned a little less than the generous wheel (12.47). Figure 5.16 shows the proportion of selectors' wheel choices over the course of the 16 periods. The probability with which the stingy wheel was selected dropped from 75% in round 1 to 25% in round 16. Conversely, the probability of selecting the fair wheel increased from 25% in round 1 to 62.5% in round

16. The probability of the generous wheel being selected remained low throughout the experiment.

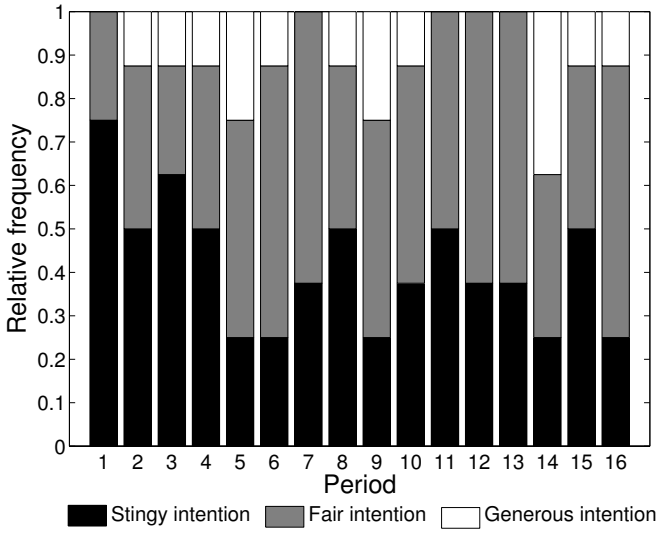


Figure 5.16: Proportion of selectors' wheel choices over the 16 periods of the game.

Overall pattern of responses Despite the monetary disincentive against adjustments, the average cost of adjustments each responder incurred was significantly different from zero ($M = 1.48$; Wilcoxon signed-rank test $Z = -2.37, p = .018$). 7 out of 8 responders made at least one adjustment in the game. In most cases (72/128), however, responders left the selector's earnings unchanged. With a ratio of negative to positive adjustments of 7:1 and a mean adjustment of -3.91 ($SD = 6.18$), responses were skewed towards subtraction. The costs incurred for adjustments did not drop in later periods, indicating that adjustments were not made strategically.

Analysis of responses: Outcomes versus intentions To evaluate the effects of intentions and outcomes in the game (Part A), we regressed responders' adjustments on dummy- coded predictors for intentions and outcomes with 'fair intention – fair outcome' as the reference category, using random intercepts for each subject (Table 5.5).[14]

With the present number of observations only the predictor for stingy intentions is significant at the conventional level ($\beta = -3.96; t = -4.03; p = .000$). Stingy intentions seem to elicit a strong negative response, with an estimated parameter that is considerably larger in absolute terms than the parameter estimate for generous intentions (2.17) and more than three times as large as the parameter estimate for a stingy outcome (-1.17). The similarity between mean adjustments in Part A and Part B (cf. Figures 5.17a and 5.17b) suggests that the presence or absence of monetary stakes had no dis-

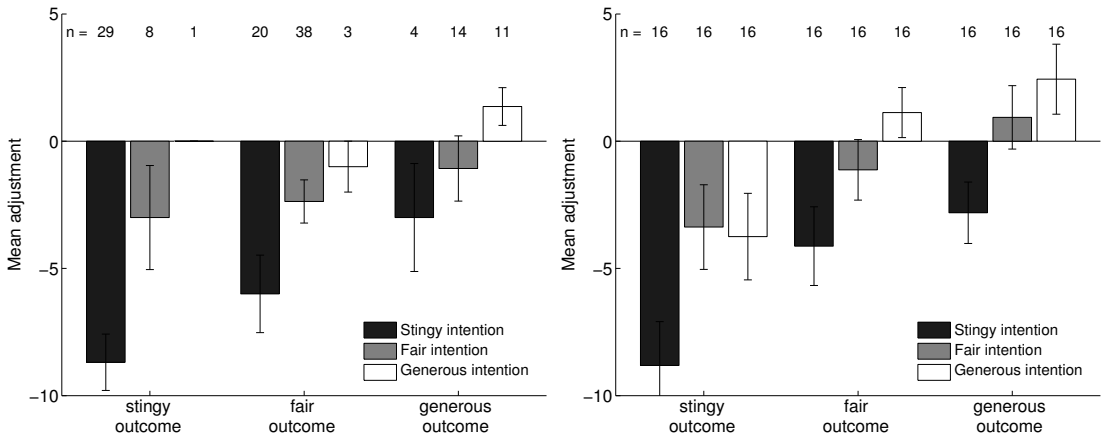
[14] A fixed intercept model was rejected in favour of random effects ($\chi^2(1) = 25.71, p = .000$, likelihood-ratio test).

Table 5.5: Regression analysis Experiment 1.

	β	t	p
Intercept	-2.48	-1.91	0.056
Stingy intention	-3.96	-4.03	0
Generous intention	2.17	1.48	0.14
Stingy outcome	-1.17	-1.11	0.269
Generous outcome	1.32	1.15	0.251

$N = 128$, $R^2 = .259$

cernible effect on participants' responses. Interestingly, however, adjustments in Part B were significantly more negative for participants who had been responders in Part A ($M = -4.67$), as opposed to selectors ($M = 0.33$; Mann-Whitney test $Z = 4.47, p = .000$). Consistent with the regression results from Part A, a repeated-measures ANOVA of Part B reveals significant main effects of intention $F(2, 30) = 11.72, p = .000$ and of outcome $F(2, 30) = 17.35, p = .000$, but no interactions $F(4, 60) = 1.16, p = .336$. The more powerful analysis for Part B indicates that on top of the effect of stingy versus fair intentions, there were significant differences in the responses to stingy versus fair outcomes $F(1, 15) = 10.97, p = .005$, as well as fair versus generous outcomes $F(1, 15) = 7.15, p = .017$.



(a) Part A **(b)** Part B

Figure 5.17: Mean adjustments by intention and outcome for (a) Part A and (b) Part B of Experiment 1 in which the wheel choice was observed. *Note*: n equals the number of observations per bar. Error bars indicate ±1 SE.

Discussion Why is it that intentions mattered more in our experiment than in Cushman et al. (2009), or in Stanca (2010), where they had little or no effect on choices? We suggest that the discrepancy could be due to a methodological difference in the elici-

tation of responses. Whereas our study used a direct-response method, both Cushman et al. and Stanca relied on the strategy method, in which participants make contingent decisions for all possible situations that may occur in the interaction. This method is advantageous in terms of data collection but may alter cognition and behaviour. Thinking hypothetically through all possible situations may induce a different style of thinking and elicit different responses than when outcomes are experienced sequentially – especially if the behaviour is moderated by emotional responses. Although in some studies the two methods do produce similar results, this seems not to hold for decisions related to punishment (Brandts & Charness, 2003), as in the present case. Moreover, the mere fact that the game was played repeatedly, albeit with different interaction partners, may have triggered a higher evaluation of intentions.

A second noteworthy result from Experiment 1 is the asymmetry in the responses to stingy and generous intentions: stingy intentions were punished fairly heavily but generous intentions were hardly rewarded. This asymmetry was also found by Cushman et al. (2009) and fits with a picture of negative reciprocity looming larger than positive reciprocity (Offerman, 2002). Offerman notes that in his study, asymmetric behavioural reciprocity correlated with an asymmetry in reported positive and negative emotions. Although both unintended and intended favourable outcomes elicited positive emotions, intended unfavourable outcomes provoked much stronger negative emotions than unintended unfavourable outcomes. The relation between emotions and responses to intentional actions is a subject for further study.

5.3.2.2 Experiment 2

The design of Experiment 2 was identical to Experiment 1 except for one feature: wheel choices were not directly observed by responders (see Figure 5.15). Instead, selectors communicated their wheel choice after having observed the outcome of the wheel. As an example, let's assume that the selector chose the selfish wheel which resulted in a selfish outcome. Both the selector and the responder are informed about the outcome (the responder, however, does not know what wheel the selector actually chose). The selector then informs the responder about the wheel that he chose. The responder, before rewarding or punishing the selector, is asked to indicate what wheel he thinks the selector actually chose.

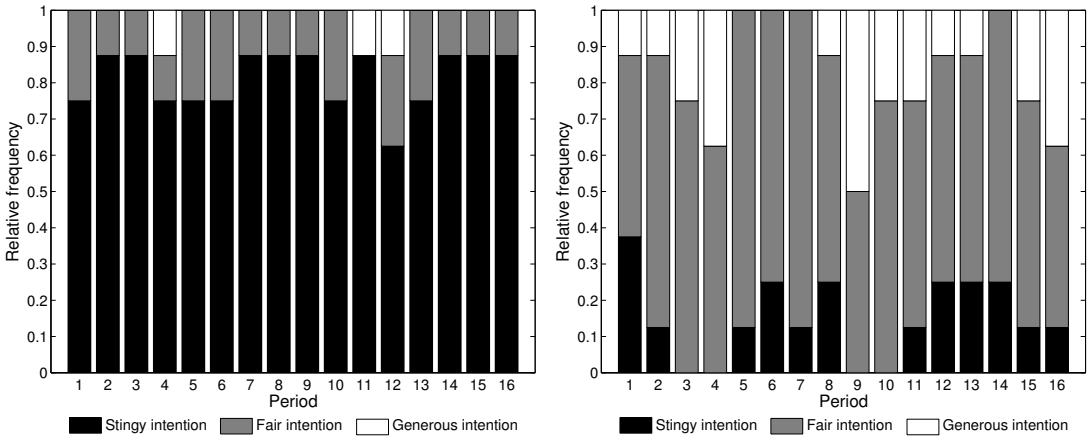
We expected selectors to make use of the possibility of deception. Both selectors and responders knew that the stated wheel choices needed not be the actual choices. Hence, we anticipated that responders would treat selectors' statements with caution. As in Experiment 1, participants first played the game for 16 times in Part A and then indicated their hypothetical responses to all nine possible combinations of stated choices and outcomes in Part B.

Method

Participants and procedure The 16 participants of Experiment 2 were recruited from the same subject pool as for Experiment 1. 13 of them were female and the median age was 22.5 years ($SD = 2.88$). Procedures were identical to Experiment 1. Experiment 1 and 2 were conducted on the same day in the same facilities. Average earnings were £11.88 ($SD = 1.20$), including a show-up fee of £5.

Results

Selectors' wheel choices and statements Figure 5.18 shows how the proportion of actual wheel choices (Figure 5.18a) and stated wheel choices (Figure 5.18b) changed over the course of the experiment. Table 5.6 shows the summarised frequencies of actual and stated wheel choices and the mean profits associated with them. Unlike in Experiment 1, selectors predominantly chose the stingy wheel ($n = 104$), followed by the fair wheel ($n = 21$) and only three choices of the generous wheel. These actual wheel choices contrast with the statements selectors made. Only in 34 of 128 cases did selectors state their true choice. In 90 cases they overstated their intentions, that is, they stated a less benign choice than they had actually made. There were 4 cases of understatements. Only in 19 cases did selectors state a stingy choice (see Figure 5.18b). Most of the time, they claimed to have chosen the fair wheel ($n = 86$) or the generous wheel ($n = 23$). In other words, selectors made heavy use of the possibility to deceive about their intentions. Selectors not only overstated their intentions, but did so in a way that takes the credibility of their statements into account. For example, a stated generous wheel choice is more credible if the outcome of the stingy wheel is generous rather than stingy. Indeed, for stingy wheel choices the proportion of stated generous



(a) Selectors' actual wheel choices **(b)** Selectors' stated wheel choices

Figure 5.18: Proportion of (a) *actual* and (b) *stated* wheel choices over the course of the experiment.

Table 5.6: Frequency and ex-post profitability of wheel choices and statements.

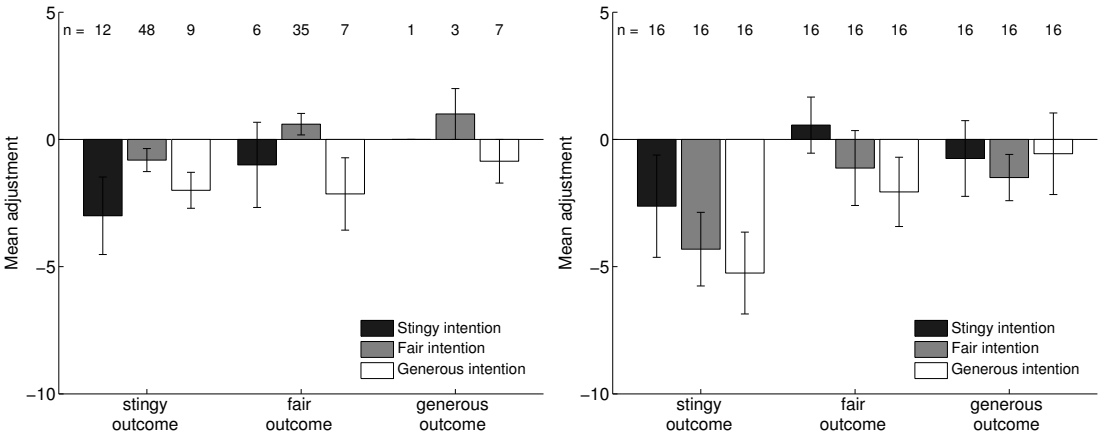
Actual wheel choice	Stated wheel choice	Mean Profit [tokens]	n
Stingy	Stingy	15	15
Stingy	Fair	18.07	69
Stingy	Generous	14.05	20
Fair	Stingy	17.67	3
Fair	Fair	14.88	17
Fair	Generous	15	1
Generous	Stingy	20	1
Generous	Fair	-	0
Generous	Generous	10	2

choices increased for a fair or generous outcome ($\chi^2(2) = 11.13, p = .004$).

Stated and believed intentions Knowing about the possibility of deception, we expected responders to treat selectors' statements with caution. In fact, according to their indicated beliefs, responders did not believe the stated intention in 62.5% of all cases. How good were responders at estimating selectors' actual wheel choices? With 57.8% correct beliefs, responders were more accurate than expected by naive random guessing (expected hit rate of 1/3) and about as accurate as expected by always going with the outcome (expected hit rate of 0.6). In fact, responders' believed intention did concur with the outcome in 57.0% of the cases. As it turns out, responders would have had the highest rate of correct beliefs (81.3%) by always assuming the stingy wheel was chosen. In summary, although responders did not take stated intentions at face value, they underestimated the proportion of stingy wheel choices.

Overall pattern of responses The mean adjustment of -0.75 ($SD = 3.30$) was considerably smaller in absolute terms than in Experiment 1 (Mann-Whitney test, $Z = -3.86; p = .000$). As in Experiment 1, responders left the selector's payoff unchanged in most cases (81/128), but a majority of responders (6 of 8) did make an adjustment at least once throughout the game. Also as in Experiment 1, adjustments were negatively skewed (30 negative vs. 17 positive adjustments) and there was no apparent drop in adjustment costs incurred in later periods, indicating non-strategic use of adjustments. There were again no clear temporal patterns of choices, statements or beliefs.

Analysis of responses: Outcomes versus stated intentions Figure 5.19a shows mean adjustments by outcome and stated intention in Part A. Comparing with Experiment 1 (Figure 5.17), we note that overall less tokens were subtracted and the effect of intentions seems to be different in the two experiments. A 3-way ANOVA confirms that there is an overall difference between experiments $F(1, 246) = 7.60, p = .006$ and



(a) Part A **(b)** Part B

Figure 5.19: Mean adjustments by intention and outcome for (a) Part A and (b) Part B of Experiment 2 in which the wheel choice was stated. *Note*: n equals the number of observations per bar. Error bars indicate ±1 SE.

an interaction between intentions and experiment $F(2, 246) = 4.14, p = .017$. No difference in the effect of outcomes is found between experiments $F(2, 246) = 0.41, p = .664$. As for Experiment 1, we regressed adjustments on dummy-coded predictors for stated intentions and outcomes, with 'stated fair intention – fair outcome' as the reference category and random intercepts for each subject (Table 5.7).[15]

Compared to statements of fair intention, both stingy and generous statements are estimated to lower adjustments by about 2 tokens, a negative effect larger than the effect of a stingy outcome (−1.28) but smaller than the effect of a stingy intention in Experiment 1 (−3.96). Generous outcomes have no significant effect on responses ($p = .601$). Looking only at stingy outcomes, the response to a stingy intention is less negative when the intention was stated as opposed to observed (Mann-Whitney test, $Z = -2.64, p = .009$). In contrast for generous outcomes, stated – unlike observed – generous intentions are punished because such statements are met with suspicion (the difference, however, is not significant $Z = -1.81, p = .126$, Mann-Whitney test). In summary, the two experiments differ in the overall level of adjustments and the effect of intentions in Part A.

The diverging effect of intentions in the two experiments becomes more pronounced in Part B (Figure 5.19b). Unlike in Experiment 1, the response patterns in Part B look different from Part A. Again, the responses of (former) selectors are significantly more negative ($M = -3.25$) than the responses of former responders ($M = -.67$; Mann-Whitney $Z = -2.97, p = .003$). However, as suggested by graphical comparison and consistent with Part A, a mixed 3-way ANOVA (Intention x Outcome x Experiment)

[15] A fixed intercept model was rejected in favour of random effects ($\chi^2(1) = 8.91, p = .003$, likelihood-ratio test).

Table 5.7: Regression results Experiment 2.

	β	t	p
Intercept	0.52	1.07	0.285
Stingy stated intention	-1.95	-2.43	0.015
Generous stated intention	-1.89	-2.37	0.018
Stingy outcome	-1.28	-2.15	0.032
Generous outcome	0.59	0.52	0.601

$N = 128$, $R^2 = .109$

reveals that the effect of intention in Part B differs between the experiments $F(2, 60) = 14.11, p = .000$ but not the effect of outcome $F(2, 60) = 1.54, p = .226$. For stingy and fair outcomes, kinder stated intentions led to *more* subtractions, whereas kinder observed intentions led to less subtractions. With stated intentions, even generous outcomes elicited subtractions. In contrast, observed generous intentions were rewarded when the outcome was generous. Once again we find that intentions influence adjustments, but the direction of the effect was reversed due to a lack of credibility of stated fair or generous intentions.

Discussion In the introduction, we stated our hypothesis that the behavioural relevance of intentions might be affected by uncertainty about intentions, and in particular, the possibility of deceptive statements. Did intentions matter less in Experiment 2, which involved stated as opposed to observed intentions? Not quite. In fact, intentions again had a larger effect than outcomes in the game. More specifically, there appears to be a positive premium for (alleged) honesty and a negative premium for (alleged) dishonesty. Stated stingy intentions were punished less than observed stingy intentions. In other words, responders seem to credit selectors for 'at least being honest' when going for the stingy wheel.

On the other hand, adjustments were less positive for stated generous intentions than for observed generous intentions. Given that these statements were seen as largely implausible, we can interpret this difference as a punishment for an attempt to deceive. This interpretation is corroborated by the fact that in Part B, former selectors subtracted more than former responders, arguably because having overstated their intentions themselves in Part A, they interpreted statements in Part B as deceptive. Note that responders could never be certain that stated generous intentions were actually a lie. Still, mean adjustments in these cases were negative, suggesting that the desire to punish dishonesty overrode concerns to avoid potentially unfair punishment (cf. Huck, 1999). The result is in line with a previous study by Brandts and Charness (2003), who found that participants punish deceptive messages about intended play in a simultaneous two-player game with pre-play messages. In contrast to our experiment, statements of intended play were made before an action was taken, not after, and deceptive statements

were revealed by observable choice. In our experiment, deception was never revealed but could only be conjectured with uncertainty.

Conclusion In addition to a concern for their own payoffs, participants in our experiments expressed a concern for intentions and honesty that was at least as large as their concern over distributional outcomes. In Experiment 1, the punishment of stingy intentions had a larger effect on adjustments than either outcomes or generous intentions. In other words, negative reciprocity loomed larger than positive reciprocity and distributional preferences per se. An asymmetry in negative and positive reciprocity has been observed before and may be connected to an asymmetry in emotional reactions. The general importance of intentions is in line with a study by Charness and Levine (2007), somewhat different from the low relevance of intentions found by Cushman et al. (2009) and at odds with the irrelevance of intentions found by Stanca (2010). To disentangle the precise design factors that influence the role of intentions is a task for future research. We suggest that the use of the strategy method and the absence of repetition may decrease the relevance of intentions.

In Experiment 2, when choices were no longer observed but merely stated, the perceived honesty of statements moderated the effect of intentions. Stated stingy intentions were punished less than observed stingy intentions, indicating an appreciation for "at least being honest". Conversely, stated generous intentions were punished, indicating a dislike for perceived attempts to deceive. Indeed, selectors routinely overstated the kindness of their intentions. Although responders did not take statements at face value, they underestimated the frequency of stingy choices. Having lied themselves, Part A selectors were particularly negative in their responses to fair or generous statements in Part B. Overall, stingy choices were more common in Experiment 2 than in Experiment 1, indicating that selectors understood the relevance of intentions and took advantage of their non-observability. The differences in response patterns between Experiment 1 and Experiment 2 underline the importance of non-outcome related factors to participants' behaviour.

We see two research questions that would be worthwhile to explore as a immediate follow up to our findings. First, we would like to see how the behaviour of both selectors and responders changes in a situation in which they repeatedly interact with each other. Intuitively, the rate of adjustments by the responder is likely to increase in such a setup since responders can directly influence their interaction partner (as opposed to the pool of interaction partners in our experiments with the repeated stranger design). On the other hand, pairs of participants might also reach a state of equilibrium (e.g. a repeated choice of the fair wheel by the selector with no adjustment by the responder). As suggested above, a repeated interaction setup might also encourage responders to focus even more strongly on the intentions of the selector. Since the selectors have only partial control over the outcome but full control over the choice of wheel, a reinforcement strategy that focuses on the selectors' intentions is more likely to yield the desired

behaviour. However, the repeated nature of the design might also highlight people's distributional concerns as they might not want to come out of the experiment with less than their partner.

It would also be interesting to see how repetition affects selectors' statements when wheel choices are not directly observed by the responders. One might expect that the high frequency of false statements observed in the stranger design would reduce when people repeatedly interact with each other. Through repeated interactions, responders are likely to be able to infer whether their partners are generally honest. As the results of the experiment have shown, responders punished allegedly dishonest statements more strongly than allegedly honest statements. Furthermore, previous research suggests that people value being seen as honest and are more likely to be dishonest when justifications for their dishonesty are readily available (see, e.g. Shalvi et al., 2011; Shalvi, Eldar, & Bereby-Meyer, 2012).

Second, the two experiments thus far have only investigated both extremes on the continuum of intention transparency: selectors choices were either always known (Experiment 1) or never (Experiment 2). It would be interesting to explore the middle ground. Just like in real life, while we normally cannot access other people's intentions directly, sometimes we find out. Imagine that John is upset about the fact that Linda forgot to send him an e-mail on his birthday. Linda, however, claims that she did indeed send out the e-mail and that it must have gotten lost somehow. The next day, John checks Linda's e-mail outbox on her computer while she is away and finds that the e-mail had never been sent. Translated into our paradigm, we would explore what happens when there is a non-zero chance that selector's actual wheel choices will be revealed. We expect that this manipulation would both lead to a decreased rate of false statements by the selectors and an increase in perceived credibility by the responders. It would be interesting to see whether both parties would benefit overall from the possibility of revealment in that they might both end up with more money on average in this setting.

5.4 Modelling the Effects of Priors on Responsibility Attributions

Developing a formal model of how differences in mental states or underlying skill level affect responsibility attributions remains a major theoretical challenge. We have seen in Chapter 4 how people's responsibility attributions to individuals in groups were systematically affected by perceived criticality and whether a person's contribution made a difference to the outcome. More precisely, people were not only sensitive to whether a person made a difference but also to how close the person was to making a difference. Thus, we saw that attributions of responsibility are closely linked to a sophisticated conception of counterfactuals.

In the discussion of the results of the knowledge experiment in Section 5.1, we have suggested that this broader kind of counterfactual reasoning provides a justification for why participants attribute responsibility to agents whose action did not make any difference to the outcome in the actual situation. The same agent's action could have made a difference to the group outcome in another possible situation. More generally, we could model the influence that different mental states have on responsibility attributions via the way in which they affect what would have happened in relevant counterfactual worlds.

Consider, for example, the asymmetry in blame attributions to skilled and unskilled players. As argued previously, when both players failed in their tasks, the counterfactual that a skilled player could have performed well is more readily available than the counterfactual that the unskilled player could have performed well. In other words, generating a counterfactual world in which a skilled player succeeded (when having failed in the actual world) requires less of a change from the actual world compared to a counterfactual world in which an unskilled player succeed (when, in fact, he failed).

If we consider the practice of responsibility attribution from a functional perspective, there is another important asymmetry. Through blaming another person we communicate that we expect them to change their behaviour in the future. If a skilled player does not perform well, we can attribute this failure to a lack of focus or concentration. These are aspects that are under the control of the player. Hence, by blaming the person we might bring about a positive behaviour change by motivating the person to be more focussed the next time. In contrast, for an unskilled person, the negative outcome is likely due to a lack of skill. Blaming an unskilled person is unlikely to result in a more positive outcome the next time especially if the person has already tried their best. Whereas effort and concentration are factors that are under a person's control, skill level can, if at all, only be changed in the long run (cf. Alicke, 2000; Weiner & Kukla, 1970). Now for a positive outcome, in contrast, there is no harm in crediting the unskilled player and signalling that he should stay motivated. The same is true for the skilled player – her performance is likely to decrease if she is not motivated to stay focused and put in all the necessary effort. How attributions of responsibility in a team are influenced by skill level and performance when the criticality of the team members is varied remains an important question for future research.

Not only a person's level of skill but also their underlying intention is closely related to counterfactual considerations. In the experiments reported in Section 5.3, we used a person's choice as a proxy for their intention. Consider, for example, that a selector chose the generous wheel but the outcome turned out to be selfish. Given the intention of the selector, the counterfactual situation in which the outcome would have been generous (or at least fair) is highly available. Thus, while it is true that the selector's action resulted in a negative outcome in this particular situation (due to the inherent noisiness of the world), the outcome was not intended and a world in which the same

choice would have resulted in a more positive outcome is readily available. Thus, the results again support the idea that people not only care about what outcome a person brought about in the actual world but also about what outcome the same action might have led to in another possible world.

Generally, whether or not an action was performed intentionally affects what we think would have happened if the situation had turned out somewhat differently. Indeed, intentions can be said to glue causes and effects together – they render the relationship between cause and effect robust to perturbations in the background conditions (Heider, 1958; Lombrozo, 2010; Woodward, 2006). Consider a boy who throws a stone toward the water and hits a swan. The intuition is strong that we blame a boy who intentionally harmed the swan more than a boy who accidentally hit the swan but actually just wanted to see his stone splash in the water. How can we capture this effect of intentionality in terms of counterfactuals? Remember the notion of *equifinality* which is at the core of Heider's (1958) distinction between impersonal and personal causation. Equifinality means that the same goal will be reached through different causal paths (cf. Figure 2.2). Accordingly, what renders the intentional action more blameworthy is the fact that it would have brought about the same negative outcome *even if* the background conditions had been different. Whereas the accidental stone throw would not have caused any harm if the swan had been in a slightly different location, an intentional throw would have adapted to this change and the same negative outcome would have obtained. Indeed, we might say that the degree to which the world would have needed to be different so that the outcome would have been undone can serve as a proxy for the strength of a person's intention. For example, imagine that there were no stones lying around next to the water but only further away from the scene. Presumably, the more it was the boy's intention to hurt the swan the further he would have gone to get a stone to do so. Hence, even in a distant counterfactual world, the negative outcome would not have been prevented.

As a first approximation, we can aim to explore the influences of mental states (or level of skill) on attributions via assuming that differences in mental states affect the prior probability that different events are likely to occur. For example, all else being equal, if a person has a strong intention of performing a certain action, the prior probability that he will do so is higher compared to when he only has a weak intention (cf. Holton, 2009). Similarly, if a person is more skilled, the conditional probability that the intended outcome will be achieved given that she acts is higher compared to a person who is less skilled.

As we have seen in Chapter 2, Brewer (1977) and later Spellman (1997) have proposed that attributed responsibility is related to the difference that an individual's contribution made to the outcome, whereby difference was defined in terms of a change in subjective degree of belief that the effect would occur. To recap, attributed responsibility (AR) is given by

$$AR_1 = p(E|C) - p(E|\neg C), \tag{5.4}$$

where $p(E|C)$ is the probability of the effect in the presence of the cause and $p(E|\neg C)$ the probability of the effect in the absence of the cause. The greater the change in subjective belief, the more responsible is C seen for the outcome.

However, there are different ways of interpreting these probabilities which lead to quite different results. Let me illustrate this via a simple causal structure in which two causes A and B bring about E in a disjunctive fashion (i.e. $E = max(A, B)$).[16] Furthermore, we assume that there are independent factors external to A and B which influence the likelihood of each event occurring. The chances of A and B occurring are $p(A) = 0.8$ and $p(B) = 0.2$, respectively. If we interpret the probabilities in Equation 5.4 as simple conditional probabilities (the probability that E happens in the presence or absence of A) we get: $AR_1 = p(E|A) - p(E|\neg A) = 1 - 0.20 = 0.8$. However, both Brewer (1977) and Spellman (1997) state that the probabilities in Equation 5.4 are to be interpreted as counterfactual probabilities (i.e. the probability that the outcome *would* have occurred in the presence of A versus the absence of A).

As we have seen in Chapter 3, Pearl (2000) has developed a formal calculus for assessing the probability of counterfactuals (see Meder, Hagmayer, & Waldmann, 2009, for empirical evidence of how people judge the probability of counterfactuals). Remember that determining the truth (in the deterministic case) or probability of a counterfactual statement involves three steps: 1) abduction, 2) intervention and 3) prediction (see Figure 5.20). We first need to condition the causal model on the observed evidence, for example, the fact that the effect E was present. That is, we need to calculate $p(A|E)$ and $p(B|E)$. Through the observation, we update all the external random factors in the model. In our example, knowing that E was present increases the chances that A and B must have been present. From knowing E (and the structure of the situation), we can infer that at least one out of A and B must have been present. Second, we consider a certain counterfactual world, for example, the world in which A was absent. We generate this counterfactual through an intervention on the variable which fixes the variable to a given value and breaks all the incoming links to that variable. Third, we evaluate the probability of the effect given our intervention $p(E_{do(\neg A)}|E)$. In words, how likely would E have been if A was absent, given that we know that E actually occurred? Interpreted in this way, we would get: $AR_2 = p(E_{do(A)}|E) - p(E_{do(\neg A)}|E) = 1 - 0.23 = 0.77$. Given the disjunctive structure, the probability of E would have been 1 if A had happened. If A had not happened, then the probability of E is determined via the (updated) probability that B would have happened (cf. Figure 5.20).

Note, however, that we have only conditioned on the fact that E was present in the abduction step. The results would be quite different, if we conditioned on A, B and E.

[16] Because A and B have no parents, $p(E|A) = p(E|do(A))$ and $p(E|B) = p(E|do(B))$ (Pearl, 2011).

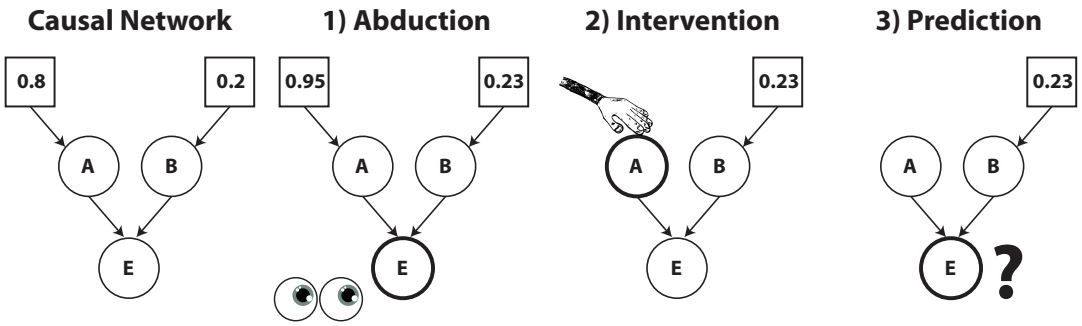


Figure 5.20: Calculation of a counterfactual's probability according to Pearl (2000). *Note*: Rectangles denote external random factors. In the abduction phase, the probabilities are updated according to Bayesian inference taking into account the disjunctive structure of the situation.

Accordingly, $AR_3 = p(E_{do(A)}|A, B, E) - p(E_{do(\neg A)}|A, B, E) = 1 - 1 = 0$. If we know that B has happened, E would have happened for sure even if A had not happened. This is due to the fact that when evaluating the counterfactual, we keep all the random factors fixed that we have updated based on our evidence (the values in the rectangles in Figure 5.20). By keeping all random factors fixed, we assure that the generated counterfactual world is maximally similar to the actually observed world.

Table 5.8 summarises the predictions of the three different versions of Brewer's (1977) model for a disjunctive and conjunctive structure. Let me briefly summarise the predictions of the different models. The AR_1 model predicts that A's responsibility is solely determined by B's prior. Varying the prior of A does not make a difference. The AR_2 model, in contrast, predicts an interaction between priors, structures and outcome. For the disjunctive structure, if the outcome is positive, A's responsibility is a function of the relative difference in priors between A and B. The greater A's prior compared to B's prior, the more responsibility A is predicted to receive for the positive outcome.[17] For negative outcomes, in contrast, A's responsibility is predicted to be full irrespective of B's prior. Conditioning on the fact that the outcome was negative in a disjunctive task fully specifies the values of the external random variables (both A and B must have been negative in this case).

The relationship between prior and outcome is reversed in the conjunctive challenge. Now, A is predicted to be held fully responsible for a positive outcome irrespective of A's or B's prior. In contrast, for negative outcomes, A's responsibility depends on the difference between priors. The lower the prior of B compared to A, the less responsible is A predicted to be for the negative outcome. Finally, the AR_3 model predicts that A's responsibility is independent of both A's and B's prior. A is predicted to receive no responsibility when the outcome is overdetermined (i.e. a positive outcome in the

[17] In a disjunctive structure with deterministic causal links, $p(B|E)$ decreases the greater $p(A)$ is compared to $p(B)$.

disjunctive structure or a negative one in the conjunctive structure) and to receive full responsibility when he is pivotal (i.e. when the outcome is negative in the disjunctive structure or positive in the conjunctive structure).

Table 5.8: Predictions of three different versions of Brewer's (1977) responsibility attribution model for a disjunctive and conjunctive structure. The top part shows attributions for a positive outcome (E) and the bottom part for a negative outcome ($\neg E$).

	Disjunctive $E = max(A, B)$	**Conjunctive** $E = max(A, B)$
$AR_1 = p(E\|A) - p(E\|\neg A)$	$1 - p(B)$	$p(B) - 0$
$AR_2 = p(E_{do(A)}\|E) - p(E_{do(\neg A)}\|E)$	$1 - p(B\|E)$	$1 - 0$
$AR_3 = p(E_{do(A)}\|A, B, E) - p(E_{do(\neg A)}\|A, B, E)$	$1 - 1$	$1 - 0$
$AR_1 = p(E\|\neg A) - p(E\|A)$	$p(B) - 1$	$0 - p(B)$
$AR_2 = p(E_{do(\neg A)}\|\neg E) - p(E_{do(A)}\|\neg E)$	$0 - 1$	$0 - p(B\|\neg E)$
$AR_3 = p(E_{do(\neg A)}\|\neg A, \neg B, \neg E) - p(E_{do(A)}\|\neg A, \neg B, \neg E)$	$0 - 1$	$0 - 0$

In sum, the three models make very different predictions about how people would attribute responsibility. Models AR_1 and AR_2 differ in terms of whether they take into account that the outcome actually occurred. In future studies, we aim to explore the predictions of these two models for situations in which participants have information about the priors of variables and know only whether or not the outcome was positive. For example, participants would know about the causal structure, how skilled two players in a team are and whether the team lost or won. We could then explore how these different factors affect the degree to which a player is held responsible for the outcome.

The models AR_2 and AR_3 differ in terms of what aspects of the situation have been observed. In AR_2 the random variables in the model are updated only based on knowing that the outcome was positive or negative. In AR_3 in contrast, the values of all variables are observed. In all the experiments reported thus far, participants did actually have full knowledge of the values of the different variables (with the exception of the deception experiment). Pearl's (2000) account of dealing with probabilistic counterfactuals does not predict any effect of differences in priors once the values of all the variables are observed. Figure 5.21 shows how much responsibility A is predicted to receive by the different models when A's prior is held fixed at $p(A) = 0.5$ and $p(B)$ is varied.

In collaboration with Noah Goodman, I am currently developing an account which extends Pearl's (2000) approach and allows for effects of priors on counterfactuals *even if* all the variables in the model have been observed. Roughly, the idea is that we allow for a small chance that the values of the variables in the counterfactual world are different from the actually observed values (cf. Lucas & Kemp, 2012; Woodward,

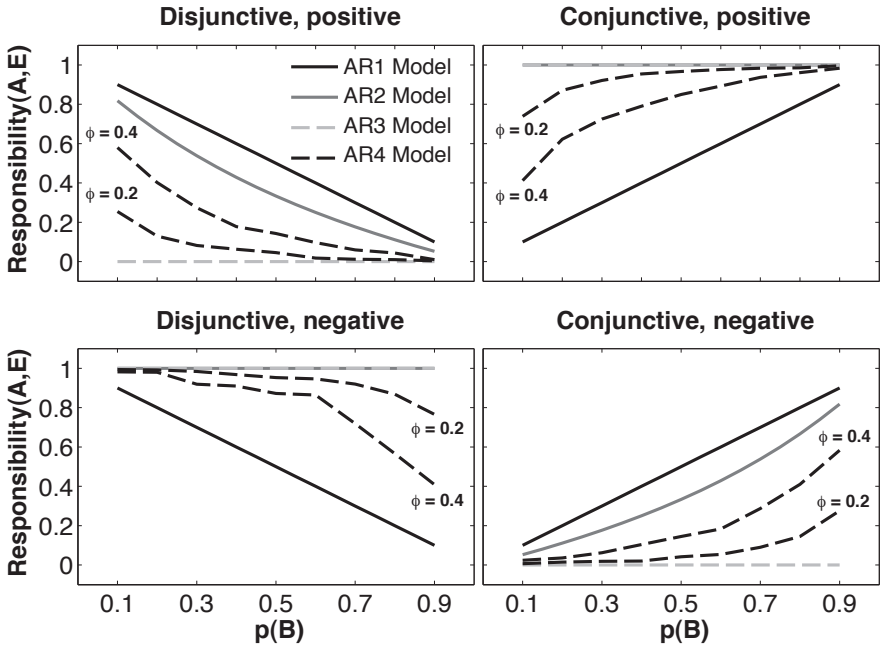


Figure 5.21: Predicted responsibility of A for E for disjunctive (left) and conjunctive (right) structures with positive (top) or negative (bottom) outcomes. $p(A)$ is fixed at 0.5 and $p(B)$ is varied on the x-axis.

2006).[18] Computationally, we condition on the observed evidence in a noisy fashion in the abduction step (see Figure 5.20). The degree of noise in the conditioning process determines to what extent the prior distribution over possible worlds is 'shifted' towards the actually observed world in order to generate the distribution over counterfactual worlds in which the effects of intervening on the variable of interest are evaluated.

If the noise parameter ϕ is fixed to 0, then the account is identical to Pearl's (2000) approach of evaluating counterfactuals. The 'distribution' of counterfactual worlds is identical to the actual world (because only the actual world is sampled). If the noise parameter ϕ is fixed to 1 (its maximum) then what happened in the actual world is ignored. The variables' values in the distribution over counterfactual worlds are determined by their priors. For intermediate values of noise, both the actual world as well as the expected world (based on the priors) influence the distribution over counterfactual worlds. Generally, the greater ϕ the more similar is the distribution over counterfactual worlds to the distribution over possible worlds based on the variable's priors. For smaller values of ϕ, the distribution over counterfactuals worlds shifts towards what happened in the actual world. According to our approach, A's responsibility for E is given by: $AR_4 = p(E_{do(A)}| \approx A, \approx B, \approx E) - p(E_{do(\neg A)}| \approx A, \approx B, \approx E)$.[19] That is, A is respon-

[18]Note that this includes variables that are independent of the variable that is intervened on in the intervention stage.

[19]The $\approx$ symbol should be interpreted as *noisy equal*. Hence, $p(E_{do(A)}| \approx B)$ reads: the probability of E when intervening on A given that B is noisily equal to its actual value.

sible to the degree that the probability of E would have been different in A's presence versus absence given that we (noisily) condition on the actual values of A, B and E.

Figure 5.21 shows the predictions of this model.[20] As can be seen, the model predicts that responsibility attributions lie between the predictions of the simple conditioning model AR_1 and the model AR_3 which is based on Pearl's (2000) counterfactual calculus assuming the values of all variables have been observed. Indeed, these two models describe the outer bounds of the AR_4 model. The setting of the noise parameter ϕ which determines the degree to which the counterfactual world is similar to the actual versus to be expected world determines how close the predictions of the AR_4 model are to the other two models. If $\phi = 0$, the predictions of AR_4 are identical to the predictions of AR_3. If $\phi = 1$, AR_4's predictions are identical to AR_1. Figure 5.21 shows two versions of the AR_4 model. If $\phi = 0.4$, the model's predictions are quite close to the predictions of the AR_1 model. In contrast, if $\phi = 0.2$, the predictions get closer to the AR_3 model.

Translated into an achievement context in which A and B represent two players and $p(A)$ and $p(B)$ the probabilities that each will be successful, AR_4 generally predicts (in line with AR_1) that A's responsibility decreases with B's prior (or skill) for disjunctive tasks and increases for conjunctive tasks. Even though we are (noisily) conditioning on having observed the values of *all* variables, the model can still predict that A is held responsible for an outcome that turned out to be overdetermined (the win in the disjunctive task and the loss in the conjunctive task).

In future research, we will test the predictions of these different models. We will explore to what extent people's attributions of responsibility are influenced by what actually happened compared to what was likely to happen a priori. The setting of the noise parameter ϕ suggests itself as an interesting parameter for exploring individual differences. As we have seen in the experiments on intended versus actual outcomes, participants differed widely in the extent to which they took the intended (i.e. the expected) versus actual outcome into account.

5.5 Conclusion

In this chapter, we have seen evidence that people are concerned about other people's mental states and their abilities when attributing responsibility. People are less likely to blame or credit an athlete when the athlete knew that their performance will not affect the team outcome anymore (Section 5.1). The degree to which people blame players also depends on their performance expectations: skilled players in a team are blamed more for the loss than unskilled players while both skilled and unskilled players receive equal credit for success (Section 5.2). Finally, we have seen that people are

[20] We assume here that for positive outcomes, both A and B succeeded in the actual world and for negative outcomes, both A and B failed. The predictions of the model are based on a sampling algorithm implemented in the probabilistic programming language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008) and are hence somewhat noisy.

not only concerned about the outcomes that other people's actions bring about but also about their underlying intentions (Section 5.3). Generally, people's attributions of responsibility (Section 5.3.1) and the extent to which they punish or reward each other (Section 5.3.2) are more strongly influenced by other's intentions than the outcomes they bring about. Modelling how differences in mental states influence attributions of responsibility remains an important theoretical challenge for future research. We suggest that a sophisticated counterfactual account provides a promising first start: to the extent that mental states shape the expectation over which world was likely to come about, we can model their influence on attributions of responsibility via showing that people do not only care about what actually happened but also about what would have happened in other possible worlds.

Chapter 6

Beyond Bayes Nets

> And now to something completely different.
>
> – **Excerpt from Monty Python's *Flying Circus***

> You can't blame gravity for falling in love.
>
> – **Albert Einstein**

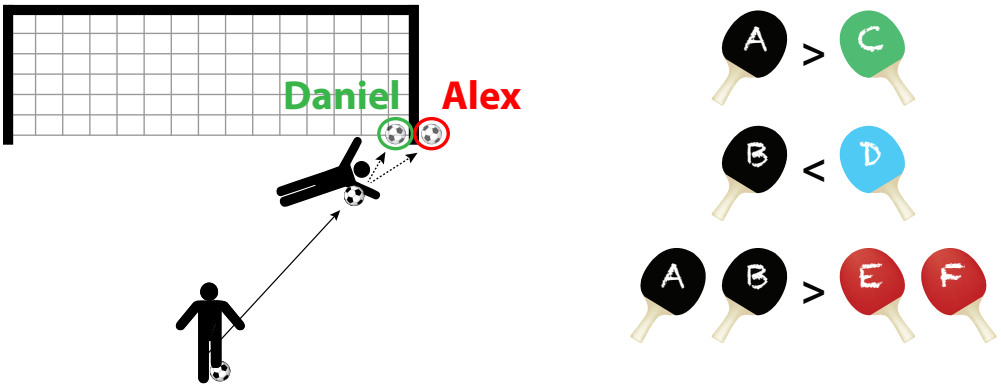
For most of the experimental paradigms we have considered so far, we were able to model the structure of the situation in terms of relatively simple causal Bayesian networks (CBN). Indeed, as argued in Chapter 3 and shown in Chapter 4, this formalism has helped us to derive and test exact predictions about people's responsibility attributions for a variety of situations which varied the size and structure of the group. Moreover, we have seen that this modelling framework suggests a very close relationship between responsibility, causality and counterfactuals.

However, CBNs also have important representational limitations. For example, they have difficulty representing situations which involve continuous space and time. Furthermore, CBNs are often constrained to express possible inferences within a particular situation but have difficulty capturing inferences that go beyond the situation at hand. Recently, cognitive scientists have become more and more aware of these limitations and have moved on to explore more sophisticated models of inference (Tenenbaum et al., 2011). These approaches agree with the basic idea of modelling cognitive functions such as the learning, reasoning, categorisation or language understanding in terms of inference over probabilistic generative models (Goodman & Stuhlmüller, 2012; Goodman, Tenenbaum, Feldman, & Griffiths, 2010; Kemp & Tenenbaum, 2009; Piantadosi, Tenenbaum, & Goodman, 2012; Stuhlmüller, Tenenbaum, & Goodman, 2010; Ullman, Goodman, & Tenenbaum, 2012). However, rather than being constrained to what can be expressed in a CBN, these novel approaches assume a richer language for building

cognitive models (see Chater, Tenenbaum, & Yuille, 2006; Goodman et al., 2008; Griffiths & Tenenbaum, 2009; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum et al., 2007).

In this chapter, we will see two examples from different domains which illustrate that it is sometimes necessary to go beyond CBNs in order to build adequate models of cognition. In the first section, I will show that people use their intuitive understanding of physics to make causal attributions in simple physical environments. As mentioned above, modelling inference in physical environments which are continuously extended in time and space is problematic within the CBN framework. In the second section, I will demonstrate how people can infer latent variables (such as a player's strength) from observations of complex patterns of evidence (such as the results of different combinations of games in a sports competition). Modelling people's inferences in this domain with CBNs is again problematic because CBNs are limited in their ability to support inferences about an object from one situation to another, potentially quite different situation.

As indicated by the Monty Python quote above, this chapter is different from the previous chapters in that 'responsibility' is not the central dependent variable anymore. However, as the following two brief examples illustrate, both people's intuitive understanding of physics and their ability to make inferences about the same object (or person) in different situations is of relevance to attributions of responsibility.



(a) Football example **(b)** Table tennis example

Figure 6.1: Two examples of domains relevant to attributions of responsibility which require inferences that go beyond what can be expressed with simple CBNs.

Imagine the following situation (see Figure 6.1a): in the final minutes of an important football game, a striker shoots the ball towards the goal, the goal keeper jumps and saves the ball. Two fans in the stadium, Daniel and Alex, observed the scene. Daniel shouts: "What a save! Thanks to the keeper we are going to win this game. If he hadn't saved that shot, it surely would have gone in." Alex replies: "What are you talking about? It was obvious that this shot would have never gone in anyhow? What a waste by the

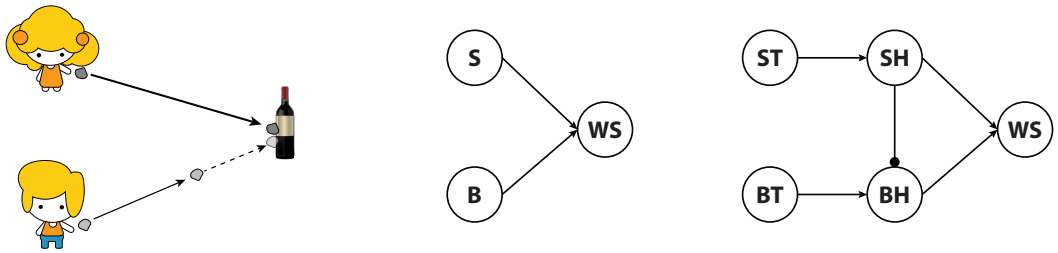
keeper to even make the effort and jump ...” As this example illustrates, whether we consider the goal keeper praiseworthy depends on what we think would have happened if the goal keeper had not saved the ball. While Daniel is full of praise for the keeper, Alex considers what happened as further evidence that the keeper is in fact useless. How can we model people's beliefs about what would have happened in the counterfactual world of interest (i.e. if the the keeper hadn't jumped)? In Section 6.1, we will see that people use their intuitive understanding of physics to simulate what would have happened if a certain event of interest had not taken place. Their certainty about the outcome of this mental simulation affects the extent to which they think the event of interest has caused or prevented the outcome.

Imagine another situation (see Figure 6.1b): Philipp observes two single player matches in a local table tennis tournament. In Game 1, player A wins against player C. In Game 2, B loses against D. After a lunch break, Philipp comes back and finds out from the score board that A and B won their doubles game against E and F. Who do you think is Philipp more likely to see responsible for the win? Here, the intuition is strong that A who won her single player match is more praiseworthy than B who lost her game. In Section 6.2, we will see how we can model this kind of reasoning.

6.1 Causal Attributions and Intuitive Physics (Gerstenberg et al., 2012)

Consider the following situation (see Figure 6.2a): “Suzy and Billy, our expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. They both pick up rocks and throw them at the bottle, but Suzy throws hers a split second before Billy. Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred, so the shattering is overdetermined. Once the bottle has shattered, however, it cannot do so again; thus the shattering of the bottle prevents the process initiated by Billy's throw from itself resulting in a shattering.” (Hall, 2004, p. 7–8)

Who, in this situation, is the actual cause of the bottle shattering? Suzy, Billy,



(a) Physical Model **(b)** Simple Model **(c)** Complex Model

Figure 6.2: Different models of the Billy and Suzy scenario.

or both? What on first sight seems like a trivial question poses a serious problem for counterfactual theories of causation. According to a simple counterfactual criterion of causation, we have to evaluate whether the bottle would have shattered if Suzy had not thrown the rock, in order to determine whether she is a cause of the bottle shattering. However, as the scenario describes, there is no counterfactual dependence between Suzy's throw and the shattering of the bottle. Even if Suzy had not thrown the rock, the bottle would still have shattered due to Billy's throw. Since the shattering of the bottle is overdetermined, none of the causes meet the strict criterion of counterfactual dependence.

Alternatively, we could apply a more relaxed criterion of counterfactual dependence according to which an event counts as an actual cause if there is a possible situation in which the effect and cause would have been counterfactually dependent (Halpern & Pearl, 2005; Hitchcock, 2001a; Yablo, 2002). If we constructed a simple model of the situation (see Figure 6.2b), we would rule both Suzy and Billy in as causes. For example, while Suzy is not pivotal for the shattering of the bottle in the actual situation, Suzy would have been pivotal if Billy had not thrown his rock. Similarly, if Suzy had not thrown, Billy would have been pivotal.

However, this simple model obscures the asymmetry that is present in the actual situation. Whereas Suzy's rock actually shattered the bottle, Billy's rock merely flew through thin air. How can a counterfactual account capture this asymmetry? One answer to this question is to say that there is no problem with the counterfactual account but merely with the way in which we have constructed the model of the situation. The simple model in Figure 6.2b does not capture the fact that there is an asymmetry between Suzy S and Billy B with respect to the shattering of the wine bottle WS. The idea would be that if we were to construct a more precise model of what happened, then the problem would be resolved (cf. Chockler & Halpern, 2004; Halpern & Pearl, 2005).

Figure 6.2c shows a slightly more complex model of the scenario. This model has two more variables than the simple model. We have unpacked both the *Suzy* and *Billy* variable of the simple model into two variables each. Accordingly, both Billy and Suzy throw their rocks (ST = Suzy throws, BT = Billy throws) and hit the bottle or not (SH = Suzy hits, BH = Billy hits). Finally, as in the simple model, there is a variable which captures whether or not the wine bottle was shattered WS. Suzy's throw causes her stone to hit the bottle which causes the bottle to shatter ($ST \rightarrow SH \rightarrow WS$). Similarly, Billy's throw causes his stone to hit the bottle and the bottle to shatter ($BT \rightarrow BH \rightarrow WS$). However, there is an additional link in the model which captures the temporal asymmetry between Suzy's and Billy's throw. Suzy hitting the bottle prevents Billy hitting the bottle (i.e. $BH = BT \wedge \neg SH$). Now that we have incorporated the asymmetry in the more complex model of the situation, we still need to make sure that the counterfactual analysis of actual causation rules in Suzy's throw ST as a cause of the bottle shattering and rules out Billy's throw BT.

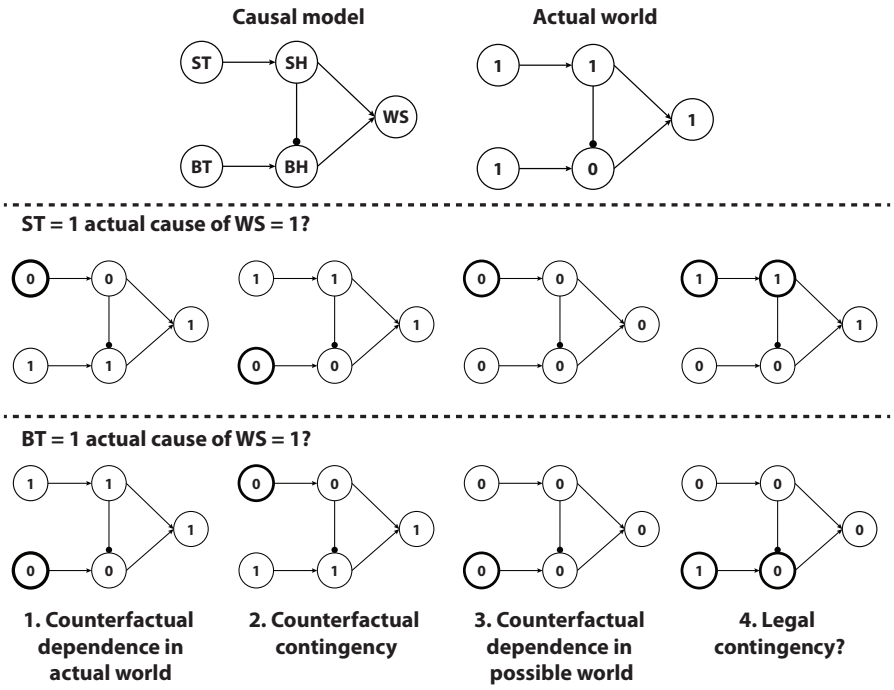


Figure 6.3: Overview of the different steps that are required to establish actual causation according to Halpern and Pearl (2005). *Note*: The variables with thicker strokes indicate what variables are considered at different steps of testing for actual causation. A variable's value of 1 means that the event took place whereas a value of 0 means that the event did not take place. For example, $ST = 0$ means that Suzy did *not* throw the stone.

Figure 6.3 summarises the different steps we have to go through in order to evaluate whether one particular event qualifies as the actual cause of another event (Halpern & Pearl, 2005). We will first evaluate whether Suzy's throw ($ST = 1$) counts as an actual cause of the bottle shattering ($WS = 1$, see middle row in Figure 6.3). In the actual world all variables are positive apart from BH which is negative (see top right of Figure 6.3). First, we check whether there is a counterfactual dependence between ST and WS in the actual world (see '1. Counterfactual dependence in actual world'). This is not the case. $WS = 1$ even if $ST = 0$. Second, we consider a counterfactual contingency in which we fix a variable that is off the active causal path connecting ST and WS (see '2. Counterfactual contingency'). We fix BT to 0 which is not on the active causal path $ST \rightarrow SH \rightarrow WS$. Third, we evaluate whether there is a counterfactual dependence between ST and WS in this situation. Indeed, this is the case (see '3. Counterfactual dependence in possible world'). Hence, $ST = 1$ does potentially count as an actual cause of $WS = 1$. Before we can conclude that $ST = 1$ is an actual cause however, we have to check that the contingency which we have created in order to test for the counterfactual dependence between ST and WS is legal.

Halpern and Pearl's (2005) definition of actual causation imposes certain constraints on what contingencies are allowed in order to evaluate whether an event counts as an

actual cause. We have already seen that one of the conditions is that we create a counterfactual contingency by only holding variables fixed at a certain value that are off the actual causal path that connects the two variables of interest. Thus, in order to evaluate whether $ST = 1$ is an actual cause of $WS = 1$, we would not be allowed to hold SH fixed at 0 for example. However, there is an additional, more subtle criterion. Setting an off-path variable to a certain value ought not interfere with the active causal path through which the cause variable of interest is connected with the effect variable. More precisely, while the setting of an off-path variable is allowed to potentially change the values of variables on the causal path, this change is not allowed to affect *the effect variable of interest*. Halpern and Pearl's (2005) way of examining that the active causal path has not been interfered with in an 'illegal way' is by checking that the original value of the effect is restored, if the variables along the active causal path are restored to their original values in the contingency in which the counterfactual dependence has been established. If this is not the case, then the considered contingency is illegal.

Hence, in a fourth step, we take the situation in which the counterfactual dependence between ST and WS was established (3) and change the variables' values on the causal path from ST to WS back to their original values to see whether WS will also change back to its original value. Indeed, this is the case (see '4. Legal contingency?'). If ST is changed back to 1 (assuming that BT had been 0) then WS is restored to its actual value 1. Hence, the considered counterfactual contingency was legal and $ST = 1$ qualifies as a cause of $WS = 1$.

Now we have to go through the same routine in order to evaluate whether $BT = 1$ counts as an actual cause of $WS = 1$. There is no counterfactual dependence between BT and WS in the actual situation (1). In a counterfactual contingency in which $ST = 0$ (2), there is a counterfactual dependence between BT and WS (3). Hence, $BT = 1$ is a potential cause of $WS = 1$. However, as it turns out, the considered counterfactual contingency is not legal (4). If we restore the values along the causal pathway from BT to WS to their original values in the contingency in which the counterfactual dependence has been established, WS will *not* change back to its original value. In fact, there is no legal contingency in which there is a counterfactual dependence between BT and WS. Thus, $BT = 1$ does not count as an actual cause of $WS = 1$.

Another way of checking whether the setting of an off-path variable is illegal is by considering whether the effect variable of interest changes when the off-path variable is changed. In Figure 6.3 we can see that when considering whether setting BT to 0 is a legal contingency to evaluate whether ST is an actual cause of the WS, the value of WS stays the same as in the actual world (see 4. Legal contingency?). In contrast, setting the off-path variable ST to 0 when considering whether BT is an actual cause of WS, changes the value of WS from 1 to 0. If the setting of an off-path variable directly changes the value of the effect variable of interest then this constitutes an illegal contingency.

This example shows how the causal model approach to actual causation can handle

cases of preemption. The question is whether this analysis of actual causation is theoretically convincing. To be fair, it does yield the desired outcome: in line with intuition, Suzy's throw comes out as an actual cause of the bottle shattering whereas Billy's throw does not. However, it feels like we need to do a lot of work to reach a conclusion which seems plainly obvious from just looking at how the actual process of stone throws unfolded over time and space (see Figure 6.2a). That being said, the general methodology of testing for actual causation applies to much more complex situations in which our intuitions might not be so clear anymore.

One way of criticising the analysis is via showing that there are situations in which this approach predicts that an event counts as an actual cause whereas people do not think so or vice versa. Livengood (2011) has shown for simple voting scenarios that Halpern and Pearl's (2005) definition yields counterintuitive predictions. Potentially even more problematic is there are structurally isomorphic situations for which the intuition is strong that the same event is an actual cause in one situation but not in the other (Hall, 2007; Hiddleston, 2005; Hitchcock, 2009). This suggests that analysing actual causation merely in terms of counterfactual dependence over causal models is not sufficient (Danks et al., in press). Recently, researchers have tried to repair the general approach by taking into account considerations about normality (Hall, 2007; Halpern, 2008; Halpern & Hitchcock, 2011, forthcoming). Inspired by research which has shown that people tend to select events that are abnormal when given a choice of possible causes (Hart & Honoré, 1959/1985; Kahneman & Miller, 1986), the idea is roughly that an event is only an actual cause when the contingency for actual causation is more *normal* than the actual contingency. When two causes are 'competing' for the status of actual causes, these accounts predict that people will be inclined to call the one the actual cause whose contingency is more normal than the other's.

The preemption example also illustrates that CBNs are limited in terms of their generality (which will be stressed more in Section 6.2). While the causal model has some generality in the sense that it allows one to say whether or not the bottle will shatter for different values of the variables, it is very much tied to what occurred in the actual situation. Indeed, in order to get the asymmetry between Billy and Suzy, we had to put in the inhibiting link from SH to BH upfront. However, we end up with a model that does not capture *in general* what happens when people throw stones at bottles. For example, we would need a new model for a situation in which Billy's throw comes slightly before Suzy's throw or in which both hit the bottle simultaneously. Thus, it feels like we are fitting the model in order to support our intuitions rather than predicting people's intuitions from the model. Arguably, the most interesting question concerns the construction of the model. How is the continuous input from the physical world translated into a psychological causal model of what has happened? So far, there is no real principled solution as to how a model ought to be constructed in order to capture the causal structure of the world (although some suggestions have been made

in Halpern & Hitchcock, 2011).

These and other problems of counterfactual accounts of causation have led philosophers to develop a concept of causation in very different terms. Indeed, there has been a longstanding divide in philosophy between two fundamentally different ways of conceptualising causality. According to *dependency accounts* of causation, what it means for C to be a cause of E is that E is in some way dependent on C. Dependence has been conceptualised in terms of regularity of succession (E is regularly succeeded by C; Hume, 1748/1975), probabilities (the presence of Cincreases the probability of E; Suppes, 1970) or counterfactuals (if C had not been present E would not have occurred; Lewis, 1973). For *process accounts*, in contrast, what it means for C to be a cause of E is that a physical quantity is transmitted along a pathway from C to E (Dowe, 2000; Salmon, 1994).

Salmon (1994) and Dowe (2000) argue that causation is fundamentally about processes. Accordingly, if one event C causes another event E to happen, there is a spatiotemporally continuous process which transmits a conserved quantity (such as physical force) from C to E. For our stone throw example, Suzy's throw is a cause because her stone transmits a force to the bottle – her stone *produced* the shattered bottle. More precisely, there is a transitive causal chain from her throw which accelerates the stone to the transmission of the stone's force onto the bottle which causes it to shatter. For Billy's throw, in contrast, there is no process that connects his stone and the bottle.

Philosophers tend to focus on developing an accurate conceptual (or metaphysical) analysis of causation, that is, understanding what causation really *is*. Psychologists, in contrast, are mostly concerned with understanding causal cognition, that is, how people perceive, learn and reason about causal relationships in the world. Some support for a process theory of causation as a plausible contender in the race of *psychological* accounts of causation comes from the domain of causal perception (see Rips, 2011; Roessler, Lerman, & Eilan, 2011; Schlottmann, 2000; Scholl & Tremoulet, 2000). In Michotte's (1946/1963) classic launching experiments, participants see animations which show the movements of two geometrical objects (diagramatically depicted in Figure 6.4).

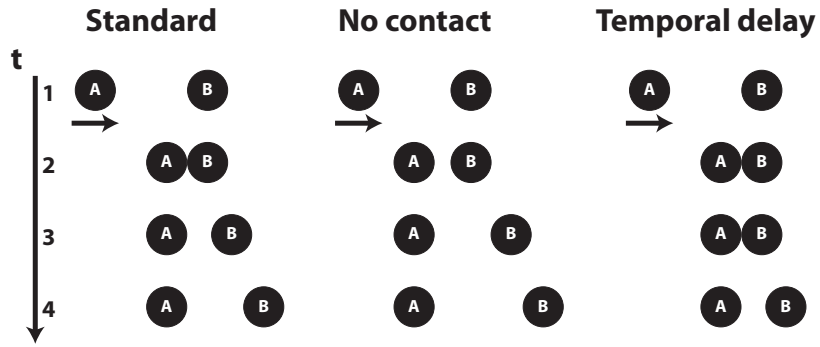


Figure 6.4: Diagramatic representation of three experimental conditions to assess people's causal perceptions. Time is from top to bottom and motion from left to right.

Participants' task is to indicate whether they perceived that one object (A) caused another (B) to move. In the *standard* condition, A moves towards B (t_1), A then stops and at the same time B starts moving (t_2–t_4). In the *no contact* condition, A stops without having made contact with B (t_2) and, at the same time, B starts moving. Finally, in the *temporal delay* condition, A makes contact with B and stops. However, there is a delay before B starts moving (t_2–t_3). While most participants indicate that they perceive A to have caused B to move in the *standard* condition, most participants do not perceive causation in the *no contact* and the *temporal delay* condition. Only if the causal process of the interaction between A and B exhibits the right tempo-spatial properties, causation is perceived (see Schlottmann & Anderson, 1993).[1]

Recently, process theories of causation have gained further popularity due to the work of Phillip Wolff and colleagues (Wolff, 2003, 2007; Wolff et al., 2010; Wolff & Song, 2003). Based on a linguistic theory of force dynamics (Talmy, 1988), Wolff (2007) developed an account of causation which explains different causal terms such as 'cause', 'enable', 'prevent' and 'despite' in terms of configurations of force vectors (see Figure 6.5). According to Wolff's account, causal relationships involve two entities: a patient (P) and an affector (A) whose forces are evaluated with respect to some endstate (E). For example, people are predicted to say that A *caused* P to reach an endstate E when P did not have a tendency towards E and A's force impacted on P in such a way that the resultant force R pointed towards E and P reached E. If P, in contrast, had a tendency towards E and A's force was concordant with P's force (i.e. both forces were parallel and pointed towards the same direction), then A is predicted to have *enabled* (or helped) P to reach E (rather than *caused*).

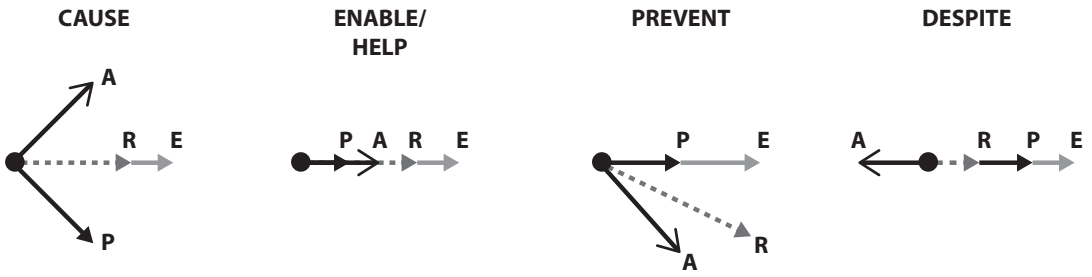


Figure 6.5: Force vector configurations that map onto different causal terms according to Wolff (2007). P = force associated with the patient, A = force associated with the affector, R = resultant force of the interaction between P and A, E = endstate.

In several experiments, Wolff (2007) presented participants with video clips in which a small boat (the patient P) raced on a pool of water on whose sides some fans (the affector A) were located. In some clips, the boat hit a small cone (the endstate E),

[1]Note, however, that causal perceptions are not discrete (i.e. either present or absent) but rather graded – some situations look more causal than others.

whereas in other clips it did not hit the cone. For each clip, participants were asked to select from a range of sentences such as "The fans *caused* the boat to hit the cone" or "The boat hit the cone *despite* the fans" Wolff's account predicted participants' modal selections of sentences for different clips very well.

Further support for the hypothesis that people conceptualise causation in terms of causal processes comes from a recent study conducted by Walsh and Sloman (2011). Participants read a series of scenarios which differed in terms of whether two events were connected merely through dependence (i.e. B would not have happened if A hadn't happened) or via a continuous causal process. Participants preferred to choose causes that influenced an effect via a continuous causal mechanism over causes that were connected with the effect through mere dependence.

However, despite their intuitive appeal, causal process theories (just like dependency theories) are plagued by some serious problems. For example, while Wolff's (2007) force dynamics model works well for interactions between physical entities, it is questionable how it can be extended to capture causal attributions in situations involving more abstract entities. One might legitimately assert that the fall of Lehman Brothers caused the financial crisis or that Tom's belief that he forgot his keys caused him to turn around and go back home. It is unclear how these causal relationships could be expressed in terms of force vectors. In contrast, they do not pose a problem for the more flexible dependency accounts. For example, according to a counterfactual account, Tom's belief qualifies as a cause of his behaviour if it is true that his behaviour would have been different had the content of his belief been different.

Furthermore, Hitchcock (1995) and Woodward (2011a) have argued that specifying causal processes in terms of a transmission of a physical quantity is insufficient. Imagine that one billiard ball A hits another billiard ball B which subsequently goes in a pocket of the table. During the collision of the two balls, A did not only transmit its velocity to B but also a bit of chalk. Of course, it is the transmission of velocity rather than chalk that did the causing. However, a causal process theory cannot trivially account for why this should be the case. Both the transmission of velocity and chalk exhibit the right spatio-temporal properties in order to qualify as causal processes. A dependency theorist can argue in terms of counterfactuals: if we had removed the chalk from A, B would still have gone in the pocket. However, if we had 'removed' the velocity from A then B would not have gone in the pocket. Thus, once a difference-making account is allowed to operate on a model with sufficiently fine granularity, it yields predictions about which properties of the causal process are causally relevant and which aren't (Woodward, 2011a).

Causation at a distance and *causation through omission* are further paradigmatic problem cases for process theories of causation. Imagine that just as Martin is about to sit down, Joe pulls away Martin's chair. In this context, it seems fine to say that Joe's pulling away the chair caused Martin to fall despite the fact that there was no continuous

process connecting Joe's action with the outcome. The situation just described falls into a class of situations called *double-prevention*. Expressed in these terms, Joe prevents the chair from preventing Martin to fall. If cases of double prevention can count as genuine causation then this is problematic for pure process accounts. Lombrozo (2010) has argued that people are especially likely to treat cases of double prevention as causation when intentional agents are involved. Thus, people should be happy to say that Joe who had the intention to make Martin fall was a cause of him falling. In contrast, people should be less willing to ascribe causal status for the same outcome to a dog bumping into the chair.

A classic example for *causation through omission* is the gardener who forgets to water the plants (Beebee, 2004; Sartorio, 2005). Did the gardener cause the plants to die? You might say that it was the hot weather rather than the gardener who caused the plants to die. However, there is certainly a counterfactual dependence between the gardener's omission to water the plants and the effect of the plants dying. Indeed, there are many situations for which people are quite happy to grant omissions causal status, for example, in situations of legal negligence. Dowe (2000) argues that omissions are not full-blown causes but only exhibit quasi-causal status (which is defined in counterfactual terms). Wolff et al. (2010) has recently developed his theory further in order to deal with causation via omission or double prevention. Interestingly, his novel account does not rely exclusively on actual forces anymore but also on what he calls *virtual forces*. Virtual forces are forces which would have been realised if some other event had not prevented their realisation. Thus, just like Dowe's (2000) account, Wolff explicitly incorporates counterfactual considerations to handle cases of omission and double prevention.

Given what we have learned thus far, one might be inclined to conclude that there is just no single conception of causation. While some problems for counterfactual accounts (e.g. preemption) are easily handled within a process framework, some problems of process theories (e.g. causation through omission) are more naturally accounted for within a difference-making framework (see McGrath, 2005). Indeed, some have argued that there are (at least) two fundamentally different concepts of causation (see Godfrey-Smith, 2010; Hall, 2004; Lombrozo, 2010; Skyrms, 1984). It has also been shown experimentally that the difference-making and process aspects of causation can be dissociated. Schlottmann and Shanks (1992) varied contiguity and contingency information using classic launching type stimuli (see Figure 6.4). They varied the temporal contiguity of cause and effect (via a delayed movement of the second block after the collision event) as well as whether an alternative event (a change of colour of the second block) was a contingent predictor of the second block's movement. They found that people's causal perception judgments were exclusively determined by the spatio-temporal contiguity of cause and effect and immune to the presence of the alternative contingent cue (i.e. the colour change of the second block). In contrast, people's judgments about the degree to which different events (such as a collision of two blocks or a change of colour

of the second block) were causally necessary for an event to occur (the movement of the second block) were indeed influenced by contingency information. Thus, while a colour change can be seen as a causally necessary condition for the movement, it does not affect people's perception of whether the movement 'looked causal' (see also Woodward, 2011a).

However, rather than committing prematurely to the view of two fundamentally different concepts of causation, we will argue for a position of unification (cf. Williamson, 2006). At least for a psychological theory of causal attribution, it seems fine that people use both process information as well as difference-making information to arrive at their judgments (see Lagnado, Waldmann, Hagmayer, & Sloman, 2007). In the spirit of Pearl (2000), we model causal attributions in terms of counterfactuals defined over probabilistic generative models. However, we agree with Wolff (2007) that people's causal knowledge is often richer than what can be expressed with CBNs. We aim to unify process and dependency accounts by showing that people have intuitive theories in the form of detailed generative models (Tenenbaum et al., 2007), and that causal judgements are made by considering counterfactuals over these intuitive theories. Here we demonstrate the superiority of our approach over existing models of causal attribution in a physical domain which involves simple collision events. We show that people use their intuitive understanding of physics to simulate possible future outcomes and that their causal attributions are a function of what actually happened and their subjective belief about what would have happened had the cause not been present.

6.1.1 Overview of experiments and model predictions

Before discussing the predictions of our model and the supporting evidence from four experiments, we first describe our experimental domain. In all experiments, participants saw the same 18 video clips which were generated by implementing the physics engine Box2D into Adobe Flash CS5. Figure 6.6 depicts a selection of the clips.[2] In each clip, there was a single collision event between a grey ball (A) and a red ball (B) which entered the scene from the right. Collisions were elastic and there was no friction. The black bars are solid walls and the red bar on the left is a gate that balls can go through. In some clips B went through the gate (e.g. clip 18) while in others it did not (e.g. clip 5). In the 18 video clips, we crossed whether ball B went through the gate given that it collided with ball A (rows in Figure 6.6: actual miss/close/hit) with whether B would go through the gate if A was not present in the scene (columns in Figure 6.6: counterfactual miss/close/hit). Participants viewed two clips for each cell (e.g. two clips for which the actual outcome was a clear hit and the counterfactual outcome was a miss – bottom left cell).

[2] All clips can be viewed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/physicsdemo.html`

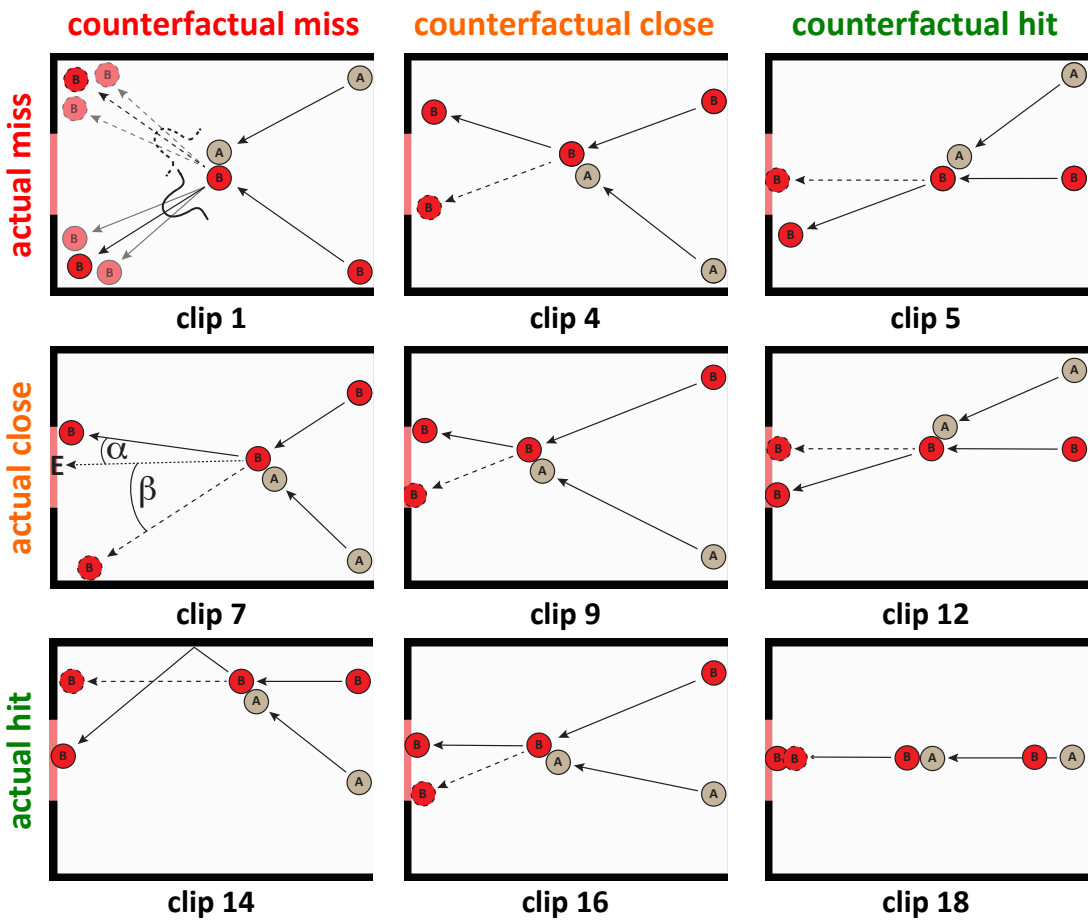


Figure 6.6: Selection of clips used in the experiment. Solid arrows = actual paths, dashed arrows = counterfactual paths. Clip 1 depicts an illustration of the Physics Simulation Model and clip 7 of the Actual Force Model. *Note*: actual miss = B clearly misses; actual close = B just misses/hits; actual hit = B clearly hits; counterfactual miss = B would have clearly missed; counterfactual close = B would have just missed/hit; counterfactual hit = B would have clearly hit.

In Experiments 1 and 2, the video clips stopped shortly after the collision event. Participants judged whether ball B will go through the gate (Experiment 1) or whether ball B would go through the gate if ball A was not present in the scene (Experiment 2). In Experiment 3, participants saw each clip played until the end and then judged to what extent ball A caused ball B to go through the gate or prevented B from going through the gate. Finally, in Experiment 4 participants chose from a set of sentences which best describes the clip they have just seen. All experiments were run online and participants were recruited via Amazon Mechanical Turk.

In order to model people's predictions of actual and counterfactual future outcomes, we developed the Physics Simulation Model (PSM) which assumes that people make use of their intuitive understanding of physics to simulate what will or what might have happened. Hamrick, Battaglia, and Tenenbaum (2011) have shown that people's stabil-

ity judgments about towers of blocks are closely in line with a noisy model of Newtonian physics. While in their model, the introduced noise captures people's uncertainty about the exact location of each block, the noise in our model captures the fact that people cannot perfectly predict the trajectory of a moving ball (see Figure 6.6, clip 1). We introduce noise via drawing different degrees of angular perturbation from a Gaussian distribution with $M = 0$ and $SD = \{1, 2, ..., 10\}$ which is then applied to B's actual velocity vector (given that it collided with A, clip 1 bottom left) or B's counterfactual velocity vector (given that A was not present in the scene, clip 1 top left) at the time of collision.

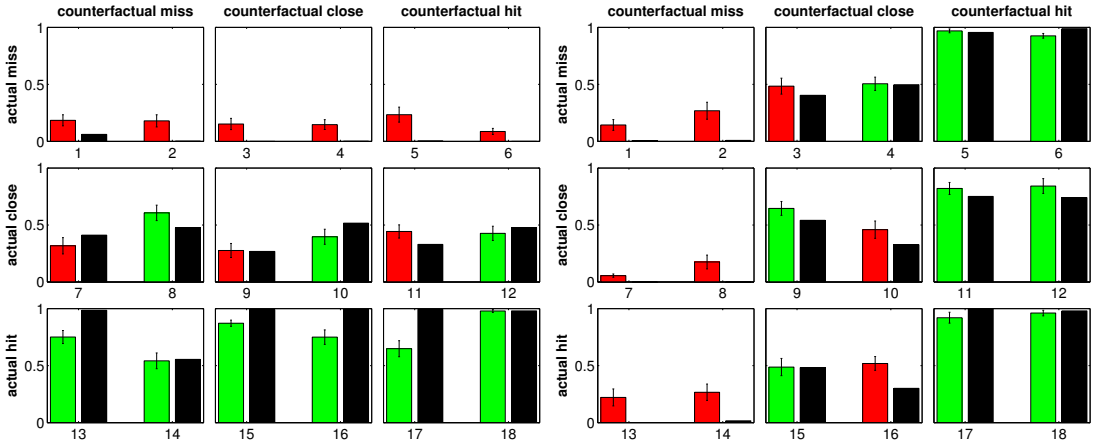
We evaluate the probability that B would go through the gate when A was present, $P(B|A)$, or absent $P(B|\neg A)$ by forward sampling from our noisy versions of Newtonian physics. For each clip and degree of noise (SD), we ran 1000 noisy repetitions of the original clip and counted the worlds in which B went through the gate given that A was present $P(B|A)$ or absent $P(B|\neg A)$.

6.1.2 Experiments 1 & 2: Intuitive physics

The aim of Experiments 1 and 2 was to evaluate how well people make use of their intuitive physical knowledge to predict actual (Experiment 1) or counterfactual (Experiment 2) future states. Participants saw 18 video clips (see Figure 6.6 for some examples) up to the point shortly after ($0.1s$) the two balls collided. After having seen a clip twice, participants answered the question: "Will the red ball go through the hole?" (Experiment 1, $N = 21$) or "Would the red ball have gone through the goal if the gray ball had not been present?" (Experiment 2, $N = 20$). Participants indicated their answers on a slider that ranged from 0 ('definitely no') to 100 ('definitely yes'). The midpoint was labeled "uncertain". After having made their judgment, participants viewed the clip until the end either with both balls being present (Experiment 1) or with ball A being removed from the scene (Experiment 2). Hence, they could check whether or not their judgment was correct. On average, it took participants 6.5 ($SD = 3.8$) minutes (Experiment 1) and 5.7 ($SD = 1.2$) minutes (Experiment 2) to complete the experiment.

6.1.2.1 Results and discussion

Participants were accurate in judging whether ball B will go through the gate (Figure 6.7a) or would have gone through the gate (Figure 6.7b) with a mean absolute difference from the deterministic physics model (which assigns a value of 100 if B goes in and 0 if B does not go in) of 28.6 ($SD = 29.9$) in Experiment 1, and 25.1 ($SD = 30.5$) in Experiment 2. Figure 6.8 shows the correlation of the PSM with participants' judgments in Experiment 1 (solid black line) and Experiment 2 (dashed black line) for different degrees of noise. While people's judgments already correlate quite well with a deterministic Newtonian physics model (degree of noise = 0°), introducing small degrees



(a) Experiment 1: Will B go in? **(b)** Experiment 2: Would B have gone in?

Figure 6.7: Participants mean judgments about whether ball B will go in (Experiment 1) or would have gone in (Experiment 2) for the 18 different clips. Green bars indicate situations in which ball B went in (Experiment 1) or would have gone in (Experiment 2) and red bars indicate situations in which ball B missed (or would have missed). Black bars are the predictions of the best fitting Gaussian perturbation model with $SD = 5°$. Error bars are ± 1 SEM.

of noise results in much higher correlations with a maximum correlation of $r = .95$ in Experiment 1 and $r = .98$ in Experiment 2 for $SD = 5°$. In sum, the results of Experiments 1 and 2 show that people are capable of mentally simulating what will happen (Experiment 1) or what would have happened (Experiment 2).

6.1.3 Experiment 3: Causation and prevention

In Experiment 3, we wanted to investigate how people use their intuitive understanding of physics to make judgments about the extent to which one event caused or prevented another event from happening. Unlike in Experiments 1 and 2, participants ($N = 22$) saw each clip played until the end. After having seen each clip twice, participants answered the question "What role did ball A play?" by moving a slider whose endpoints were labeled with "it prevented B from going through the hole" and "it caused B to go through the hole". The midpoint was labeled 'neither'. The slider ranged from -100 ('prevented') to 100 ('caused'). Participants were instructed that they could use intermediate values on the slider to indicate that ball A *somewhat* caused or prevented B. On average, it took participants 8.2 ($SD = 2.5$) minutes to complete the experiment.

6.1.3.1 Model predictions

Physics Simulation Model According to the PSM, people arrive at their cause and prevention judgments by comparing what actually happened with what they think would have happened if the cause event had not taken place. More specifically, our model

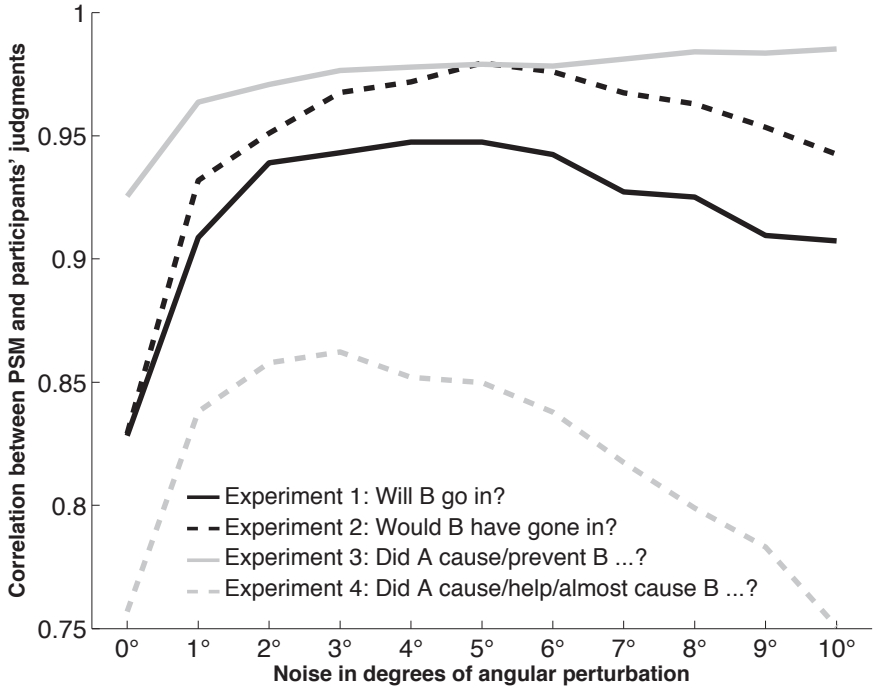


Figure 6.8: Correlation of the Physics Simulation Model with people's judgments in all four experiments for different degrees of noise.

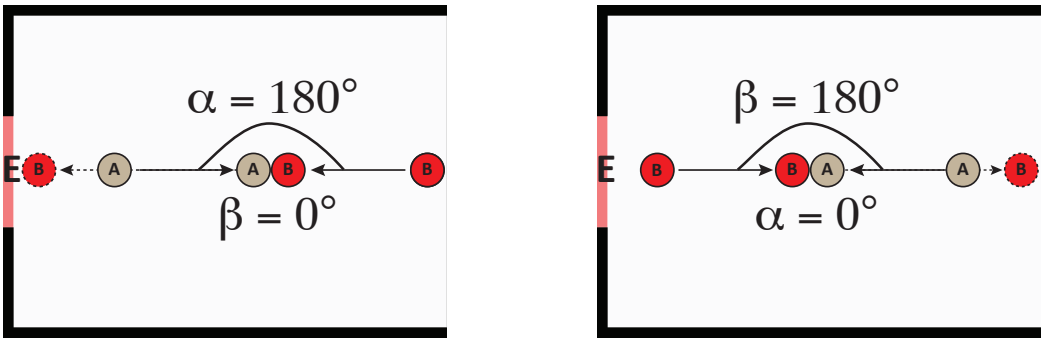
predicts that people compare $P(B|A)$, the probability that ball B will go through the gate given that it collided with ball A, with $P(B|\neg A)$, the probability that B would have gone through the gate if A had not been present in the scene. Since participants in Experiment 3 watch the clips until the end, the value of $P(B|A)$ is certain: it is either 1 when B goes through the gate or 0 when B misses the gate. In order to determine $P(B|\neg A)$, the PSM assumes that people use their confidence in the result of their mental simulation of what would have happened had A not been present.

In general, if $P(B|A)-P(B|\neg A)$ is negative, participants should say that A prevented B from going through the gate. Intuitively, if it was obvious that B would have gone in had A not been present (i.e. $P(B|\neg A)$ is high) but B misses the gate as a result of colliding with A (i.e. $P(B|A) = 0$), A should be judged to have prevented B from going through the gate. Similarly, if the difference is positive, participants should say that A caused B to go through the gate. If the chance that B would have gone through the goal without A was low but, as a result of colliding with A, B goes through the gate, A should be judged to have caused B to go through the gate. Clip 1 in Figure 6.6 shows an example for which our model predicts that participants will say that A neither caused nor prevented B. $P(B|A)$ is 0 since B does not go through the gate. However, $P(B|\neg A)$ is also close to 0 since it is clear that B would have missed the gate anyhow. Similarly, people should also say that A neither caused nor prevented B when B does in fact go in but it was clear that it would have gone in even if A had not been present

(see clip 18).

Actual Force Model The Actual Force Model (AFM) is our best attempt to apply Wolff's (2007) force dynamics model to our task.[3] According to the AFM, participants' cause and prevention judgments are a direct result of the physical forces which are present at the time of collision.

Clip 7 in Figure 6.6 illustrates how the AFM works. First, a goal vector (dotted arrow) is drawn from ball B's location at the time of collision to an end state (E), which we defined to be in the centre of the gate. Second, the angle α between the velocity vector that ball B had shortly *after* the collision with A (solid arrow) and the goal vector as well as the angle β between the velocity vector that ball B had shortly *before* colliding with A (dashed arrow) are determined. Third, the model predicts people's cause and prevention judgments via comparison of α and β. In general, if ball B goes in and $\beta - \alpha$ is greater than 0, the model predicts people will say that A caused B. Conversely, if ball B does not go in and $\beta - \alpha$ is smaller than 0, the model predicts people will say A prevented B. For situations in which $\beta - \alpha$ is greater than 0 but B does not go in or $\beta - \alpha$ is smaller than 0 but B does go in, we fix the model prediction to 0. This constraint prevents the model from predicting, for example, that people will say "A caused B" when B missed the gate.



(a) Maximum prevention: B initially moves toward the gate (E).

(b) Maximum causation: B initially moves away from the gate (E).

Figure 6.9: Diagrammatic representations of the two extreme cases for the AFM. (a) B initially moves along the ideal path (i.e. $\beta = 0°$), A collides with B and makes B go in the opposite direction from the gate (i.e. $\alpha = 180°$), (b) B initially moves in the opposite direction from E (i.e. $\beta = 180°$), A collides with B and makes B go through the gate (i.e. $\alpha = 0°$).

Figure 6.9 shows two cases which are at the opposite ends of the continuum from maximum prevention 6.9a to maximum causation 6.9b. Note that our implementation differs from Wolff's (2007) model in that A is predicted to have prevented B even when

[3]While the force dynamics model only makes predictions about which out of several sentences participants will choose to describe a situation, the AFM makes quantitative predictions about the extent to which an event is seen as causal/preventive.

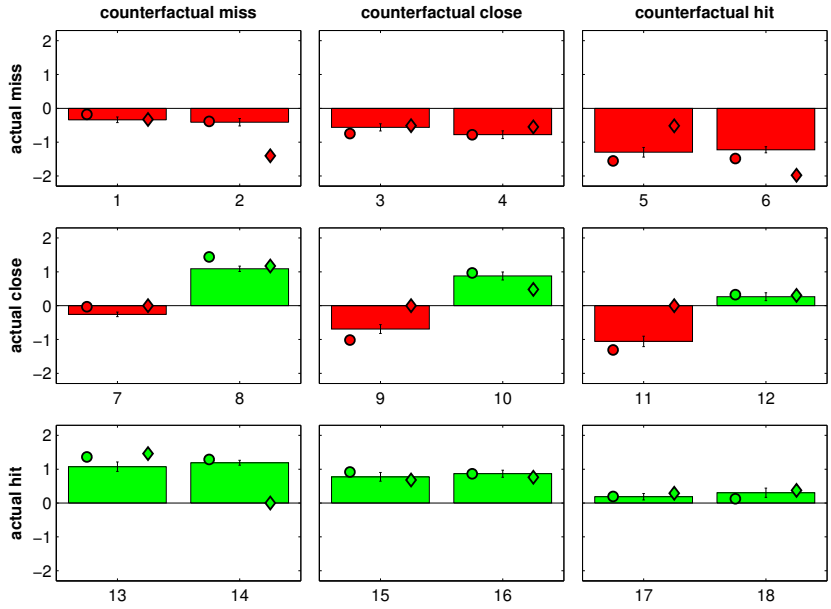


Figure 6.10: Z-scored mean cause (green) and prevention ratings (red) for the different clips denoted on the x-axes. ∘ = predictions of the Physics Simulation Model ($r = .99$), ⋄ = predictions of the Actual Force Model ($r = .77$). Error bars are ± 1 *SEM*.

B's initial force was *not* directed towards the endstate. As long as B misses the gate and α is bigger than β, the AFM predicts that A prevented B. Similarly, A is sometimes predicted to have caused B even though B did have an initial tendency towards E. When B goes through the gate and α is smaller than β, A is predicted to have caused B to go through the gate. Thus, the AFM predicts causal attributions as a function of (i) the actual outcome (whether B went in or not) and (ii) via a comparison of the angle between B's velocity vector and the ideal vector shortly before and after the collision.

6.1.3.2 Results and discussion

Figure 6.10 shows participants' mean cause and prevention judgments for the 18 different clips together with the predictions of the PSM and the AFM. For the particular implementation of the PSM depicted in Figure 6.10, we directly used participants' judgments from Experiment 2 in which they indicated whether ball B would have gone through the gate if ball A had not been present as the values for $P(B|\neg A)$. For example, in clip 5 the ball misses the gate (hence $P(B|A) = 0$) and participants' average confidence rating from Experiment 2 of whether B would have gone through in the absence of A is 97% (hence $P(B|\neg A) = .97$). Thus the PSM predicts that participants will indicate that A strongly prevented B in this clip, because $P(B|A) - P(B|\neg A) = 0 - 0.97 = -.97$ is close to the minimum of -1 in this situation.[4]

Overall, the PSM predicts participants' cause and prevention ratings very well with

[4]Note that both model predictions and participants' ratings were z-scored in Figure 6.10.

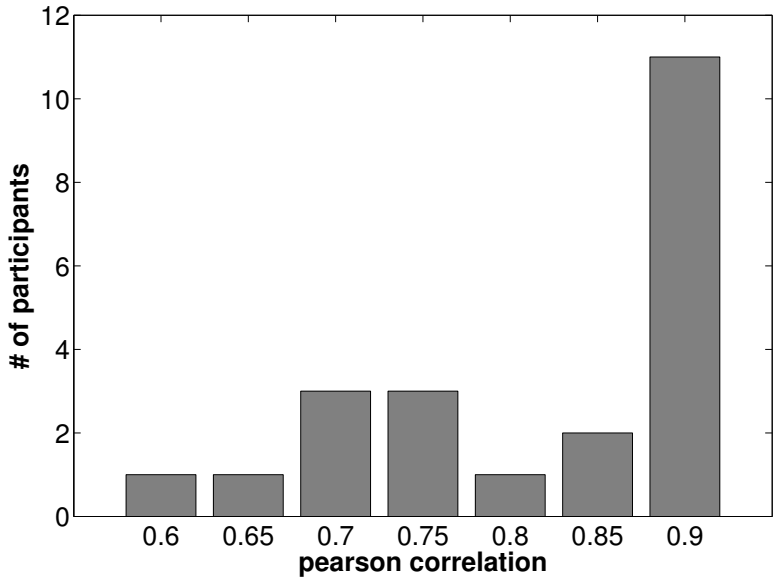


Figure 6.11: Histogram of individual participants' correlations with the PSM.

$r = .99$ and $RMSE = 0.02$. A high median correlation across participants of $r = .88$ with a minimum of $r = .61$ and a maximum of $r = .95$ demonstrates that the good performance of the PSM is not due to a mere aggregation effect (see Figure 6.11). The PSM achieves its high predictive accuracy without the need for any free parameters. We directly used participants' judgments from Experiment 2 to determine the value of $P(B|\neg A)$ for each clip (and $P(B|A)$ is fixed by the actual outcome). Figure 6.8 shows that participants' judgments also correlate highly with the PSM when we generate $P(B|\neg A)$ through the noisy simulations of Newtonian physics as described above. The AFM, in contrast, does not predict participants' judgments equally well with a correlation of $r = .77$ and $RMSE = 0.44$. While the AFM predicts people's judgments for many of the clips, there are a number of clips for which its predictions are inaccurate (most notably: clips 2, 5, 9, 11 and 14).

Interestingly, people's cause and prevention judgments were not affected by the closeness of the actual outcome. That is, participants' cause ratings did not differ between situations in which B just went through the gate (clips 8, 10, 12: $M = .51,\ SD = .40$) compared to situations in which B clearly went through (clips 13–18: $M = .49,\ SD = .42$). Similarly, prevention judgments were not different between situations in which B just missed (clips 7, 9, 11: $M = -.41,\ SD = .43$) and situations in which B clearly missed (clips 1–6: $M = -.47,\ SD = .44$).

In sum, people's cause and prevention judgments were very well predicted by the PSM. In order to judge whether ball A caused or prevented ball B, participants appear to compare what actually happened with what they think would have happened had A not been present. A very high correlation is achieved without the need for any free parameters in the model. The AFM which assumes that people arrive at their

judgments via comparing instantaneous force vectors – rather than a mental simulation of the full physical dynamics – cannot capture people's judgments equally well. Clip 14 (see Figure 6.6) gives an example in which the AFM gets it wrong. While participants indicate that A caused B to go through the gate (see Figure 6.10), the AFM model cannot predict this. In this situation, the angle between the velocity vector of B shortly after the collision and the goal vector α is greater than the angle between the velocity vector of B shortly before the collision and the goal vector β. Hence, the model predicts that A is preventing B but since B does in fact go in, the model's prediction is fixed to 0. In defence of the AFM, it could be argued that clip 14 is better thought of as a causal chain in which A causes B to hit the wall which then causes B to go in. Whether participants would count the static wall as a cause of B going through the gate is an empirical question. In any case, the other problematic clips mentioned above remain. Each of these clips only involves a single interaction. Indeed, removing clip 14 from the set only leads to a marginal increase of AFM's overall correlation from $r = .77$ to $r = .81$.

6.1.4 Experiment 4: Almost caused/prevented

The results of Experiment 3 show that people's cause and prevention judgments are only influenced by their degree of belief about whether the event of interest would have happened without the cause being present and not influenced by how close the outcome actually was. However, often the closeness with which something happened clearly matters to us, such as when we almost missed a flight to Japan or only just made it in time for our PhD viva (cf. Kahneman & Varey, 1990).

As mentioned above, one of the appeals of process accounts is that they acknowledge the semantic richness of the concept of causation by making predictions about which out of several causal verbs people will choose to describe a particular situation. In this experiment, we will demonstrate that our framework is not only capable of capturing the difference between causal verbs such as *caused* or *helped* but also predicts when people make use of intrinsically counterfactual concepts such as *almost caused* or *almost prevented*. Current process accounts (e.g. Wolff, 2007) cannot make predictions in these situations as they aim to analyse causality without making reference to counterfactuals.

In Experiment 4, participants ($N = 41$) had to select one out of seven sentences which described the situation best. The sentences were:

1. A *caused* B to go in the hole.
2. A *helped* B to go in the hole.
3. A *almost caused* B to go in the hole.
4. A *prevented* B from going in the hole.
5. A *helped to prevent* B from going in the hole.

6. A *almost prevented* B from going in the hole.

7. A *had no significant effect* on whether B went in the hole or not.

6.1.4.1 Model predictions

Table 6.1 gives an overview of the model predictions which are a function of the actual outcome, that is, whether or not B went in, and the probabilities $P(B|\neg A)$, $P(almost\ B|A)$ and $P(almost\ \neg B|A)$.

For $P(B|\neg A)$ we can again use participants' judgments from Experiment 2 or the predictions of the PSM. The model's predictions for *caused* and *prevented* are identical to the predictions in Experiment 3. Unlike Wolff's (2007) model which predicts that the difference between *caused* and *helped* is a function of the concordance of the patient's and affector's forces (see Figure 6.5), our model predicts that the difference is an epistemic one. People are predicted to select *helped* when B went in and they are uncertain about what would have happened had A not been present. When B went in *and* they are sure that it would have missed, participants are predicted to say that A *caused* B to go in. Finally, when participants are sure that ball B would have gone in anyhow, our model predicts that people will select that A had no significant effect on whether or not B went in.

The same epistemic distinction as a function of the certainty about the counterfactual outcome holds for *prevented*, *helped to prevent* and *had no significant effect* in situations in which ball B missed. When participants are sure that B would have gone in, they are predicted to say that A *prevented* B. When they are unsure about whether B would have gone in they should say that A *helped to prevent* B. Lastly, when they are sure that B would have missed the gate anyhow, they are predicted to select that A *had no significant effect* on B.

Table 6.1: Predicted probability of choosing different sentences in Experiment 4.

predicted selection	**actual outcome**	**probability**
(1) caused	hit	$1 - P(B\|\neg A)$
(2) helped	hit	$1 - abs(P(B\|\neg A) - 0.5)^{\dagger}$
(3) almost caused	miss	$p(almost\ B\|A) - P(B\|\neg A)^{\ddagger}$
(4) prevented	miss	$p(B\|\neg A)$
(5) helped to prevent	miss	$1 - abs(p(B\|\neg A) - 0.5)^{\dagger}$
(6) almost prevented	hit	$p(almost\ \neg B\|A) - (1 - P(B\|\neg A))^{\ddagger}$
(7) no effect	hit/miss	$1 - max((1), \ldots, (6))$

†rescaled to range from 0 to 1, ‡fixed to 0 for negative values.

In order to predict when people select *almost caused* or *almost prevented*, we first have to define the probabilities $P(almost\ B|A)$ and $P(almost\ \neg B|A)$. These probabilities express the closeness of an alternative counterfactual outcome to the actual outcome. One way to get at the probabilities would be to ask for participants' judgments of how closely B hit or missed the gate. However, here we used a variation of the PSM to generate these probabilities. For each clip we ran a set of 100 x 10 noisy simulations for different noise levels from $SD = 1°$ to $5°$, whereby the noise was again introduced at the time of collision. If the outcome in the noisy simulations was different from the original outcome in *any* of the ten repetitions in each of the 100 simulated worlds, we counted this as a positive instance. If the outcome in all ten repetitions was the same as the original outcome, we counted this as a negative instance.

For example, a value of $P(almost\ \neg B|A) = .87$ in a situation in which B goes through the gate in the original clip, means that in 87 out of the 100 generated worlds, the ball did not go through the gate *in at least one* of the ten repetitions of each of the worlds. For the remaining 13 worlds, the ball did go in *for all ten* noisy repetitions. Intuitively, in situations in which the outcome was close, the chances that the outcome in the noisy simulation will be different from the outcome in the original clip in at least one out of ten repetitions are high. However, if ball B clearly missed, for example, it is unlikely that there will be a noisy simulation in which the introduced angular perturbation is sufficient to make B go in.

The model predicts that people will select *almost caused* when B just missed (which means that $P(almost\ B|A)$ is high) and the probability that it would have gone in given that A was absent $P(B|\neg A)$ is low. People should select *almost prevented* when B just went in ($P(almost\ \neg B|A)$ is high), and when it was clear that B would have gone in had A been absent ($P(B|\neg A)$ is high). Finally, if none of these calculations result in a high value, people are predicted to select that A had *no significant effect* on whether B went through the gate.

The model predictions for the 18 different clips can be seen in Figure 6.12. We used Luce's (1959) choice rule to transform the different probabilities into predictions about the frequencies with which the different sentences will be selected. The model predicts that the modal choice in situations in which B does not go in changes from *prevented* for clips in which it was clear that B would have gone in (clips 5 & 6) to *helped to prevent* in situations in which it was unclear whether B would have gone in (clips 3 & 4). The same switch of the modal response as a function about the certainty of what would have happened is predicted for *caused* (clips 13 & 14) and *helped* (clips 15 & 16). If there was little uncertainty about the counterfactual outcome and it matches the actual outcome, people are predicted to select *had no effect* (clips 1, 2, 17, 18). For clip 7 in which B just misses, people are predicted to select *almost caused* since it was clear that B would have missed but for A. Conversely, for clip 12, the model predicts that people will select *almost prevented*: B just goes in and it was clear that it would have gone in without A

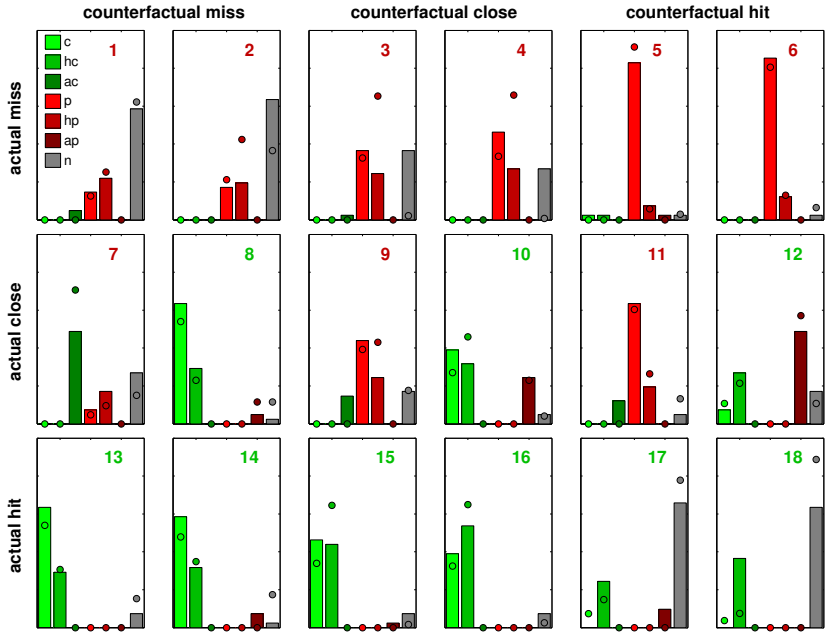


Figure 6.12: Frequencies with which different sentences were selected in Experiment 4 (bars) and predictions by the Physics Simulation Model (circles), $r = .86$. The colour of the clip number indicates if the ball went in (green) or not (red). *Note*: c = caused, hc = helped (to cause), ac = almost caused, p = prevented, hp = helped to prevent, ap = almost prevented, n = no significant effect.

being present.

6.1.4.2 Results and discussion

The model predicts the frequencies with which participants select the different sentences very well, $r = .86$ (see Figure 6.12). Figure 6.8 shows the correlation when we generate $P(B|\neg A)$ through noisily perturbing the vector rather than taking participants' ratings from Experiment 2. It predicts the modal choice correctly in 12 out of 18 clips. While participants' modal response does not change between clips 5 & 6 and clips 3 & 4 as predicted by the model, the proportion of *helped to prevent* selections clearly increases. A similar shift is observed between clips 13 & 14 and 15 & 16 for which participants' selection of *helped* increases as a function of the uncertainty over what would have happened.

As predicted by the model, participants' modal response in clip 7 is *almost caused* and in clip 12 *almost prevented*. The variance in responses within a clip is greater for the clips in which the actual outcome was close (middle row) compared to when it clearly missed (top row) or clearly hit (bottom row). For example, in clip 10 in which B just goes in and the counterfactual outcome is close the majority of participants selected *caused* or *helped* while a minority of participants selected *almost prevented*. This pattern closely matches the predictions of our model. Whether a participant is

expected to select *caused* or *almost prevented* depends on the participant's subjective belief about the counterfactual outcome. If a participant thought that B would have missed she will say A *caused* or *helped* it. However, if a participant thought that B would have gone in but for A he will select *almost prevented* because B barely went in.

The close fit between our model predictions and participants' selection of sentences demonstrates that the PSM is capable of capturing some of the richness of people's causal vocabulary. Our model not only allows to distinguish cases of *causing/preventing* from *helping* but also accurately predicts people's *almost caused/prevented* judgments.

6.1.5 General discussion

In this section, we developed a framework for understanding causal attributions that aims to break the longstanding dichotomy between process accounts and dependency accounts of causation. We showed that people's quantitative cause and prevention judgments (Experiment 3) as well as people's use of different causal verbs (Experiment 4) are well predicted by assuming that people compare what actually happened when the cause was present, with what they think would have happened in the absence of the cause. We provided evidence that people use their intuitive understanding of physics to simulate possible outcomes (Experiments 1 & 2). Our model retains the generality of dependency accounts while the use of a generative model based on Newtonian physics allows us to capture some of the richness of people's concept of causation.

According to our account, causal attributions are subjective and model-dependent. Two observers with a different understanding of the underlying generative model (or indeed, different motivations as illustrated via the goalkeeper example above) are predicted to reach different causal verdicts for the same clip when their beliefs about what would have happened in the absence of the cause event differ (see Teigen, Kanten, & Terum, 2011, for evidence that people's counterfactual judgments can be biased towards extremes). The noisy Newtonian physics model predicted participants' judgments well in our experiments. However, we are not committed to this particular generative model – indeed, our account predicts that the ways in which people's intuitive understanding of physics is biased will be mirrored in their causal attributions. While people's judgments in physical domains have been well accounted for by noisy Newtonian approximations (e.g Hamrick et al., 2011; Smith & Vul, 2012) there are well-known examples for situations in which people's intuitive understanding of physics is inaccurate (McCloskey, Caramazza, & Green, 1980; McCloskey, Washburn, & Felch, 1983; White, 2007).

For example, McCloskey et al. (1983) have shown that many people falsely believe that a moving object will (if dropped) fall down in a straight vertical line although, in fact, the object will fall forward in a parabolic arc. Imagine a situation in which two planes A and B fly over a house (see Figure 6.13). You know that at t_1 one of the two planes dropped off a bomb. However, you cannot see which one it was due to a thick

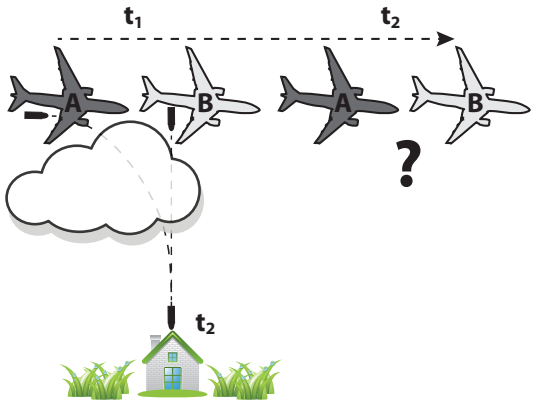


Figure 6.13: Two planes flying over a house and dropping off a bomb. Which plane caused the house to explode?

cloud. At t_2, the house explodes. Which of the two planes, A or B caused the house to explode? Which of the two pilots is responsible for the destruction of the house? Our account predicts that people's intuitive understanding of physics will determine their causal attributions. Hence, people who think that objects fall down in a straight line will attribute causality to plane B which was exactly above the house at t_1. In contrast, people who believe that objects fall down in an arc will say that plane A caused the explosion of the house.

The close match of people's judgments with the predictions of the physics simulation model of whether ball B will go through the gate (Experiment 1) or would have gone through the gate (Experiment 2), support the assumption that people are capable of simulating future or counterfactual outcomes. However, the evidence thus far is only indirect. Currently, we are running an eye-tracking experiment to get more direct evidence about whether people do indeed simulate and, if they do, how. In a recent study, Crespi, Robino, Silva, and de'Sperati (2012) had participants watch video clips of snooker players' shots in which the last part of the ball's trajectory was occluded. Participants' task was to judge whether a skittle placed in the centre of the pool table will be hit by the ball or not. They found that experts' judgments were more accurate than novices'.

More interestingly, an analysis of participants' eye-movements revealed striking differences between experts and novices. Whereas novices tended to simulate the trajectory of the (now occluded) ball with their eye-gaze in an analogue fashion, experts focused on specific diagnostic areas such as the point at which a ball hit the cushion. It will be interesting to explore whether participants in our experiments simulate (if they do) via continuous pursuit movements of the eyes or via directly jumping to the point of interest (i.e. the gate or the wall). Moreover, as Crespi et al.'s (2012) results suggest, it might be possible to observe a shift of simulation strategies over the course of an experiment. With increased experience of the experimental domain, participants might switch from an analogue to a more rule-based way of simulating (see also Hegarty, 1992,

2004; Hegarty, Just, & Morrison, 1988).[5]

While our framework shares some of the key insights of Wolff's (2007) force dynamic account, such as the need for a richer specification of people's causal representations that goes beyond simple causal Bayes nets, our proposals are different in critical respects. Most importantly, our accounts differ in the role that counterfactuals play. Wolff (2007) aims to reduce causal attributions to configurations of force vectors and argues that these force representations (which are primary) can then be used for the simulation of counterfactual outcomes (which are secondary). Our account, in contrast, does not try to explain causal attributions in terms of non-causal representations but postulates that causal attributions are intimately linked with the simulation of counterfactuals (cf. Woodward, 2011b). Hence, we claim that in order to say whether A caused B, it is necessary to consider what would have happened to B in the absence of A and not sufficient to only consider what actual (or virtual) forces were present at the time of interaction between A and B.

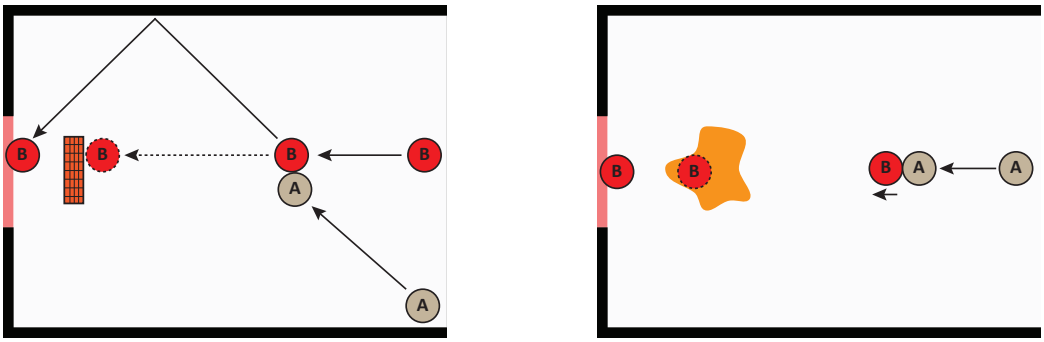
The importance of relying on counterfactuals can be nicely illustrated via the example shown in Figure 6.14a. We have amended the basic setup of our experiments described above by placing a brick in front of the gate. Remember that in Wolff's account, the patient's initial tendency is crucial. A tendency is defined in terms of the direction of a force at the time of interaction. A patient has a tendency towards the endstate if the patient's force points towards the endstate at the time of interaction. According to this definition, ball B had a tendency towards the endstate at the time of collision with A in Figure 6.14a. Since B had a tendency towards the endstate (E), it reached E and B's and A's forces were not concordant, none of the force diagrams fits and Wolff's (2007) accounts yields no prediction for this case. Although we haven't run the experiment yet, I am fairly confident that people will say that A caused B to go through the gate in this situation. Our physics simulation model predicts this: B does in fact go through the gate and it would not have if A hadn't been present.

Our account differs from Wolff's in that we give a counterfactual interpretation to the patient's tendency. In our account, a patient's tendency is captured by whether or not people think (using their intuitive understanding of physics) that the patient will reach the endstate. Examples like the one just described illustrate that defining tendency merely in terms of the direction of the patient's force at the time of collision is insufficient.

Figure 6.14b shows another interesting test case for which the predictions of our account are different from Wolff's force dynamic account. In this situation, there is a puddle of beer in front of the gate.[6] B was already moving toward the gate. A bumped into B and speeded B up. As a result, B went through the gate. The configuration of

[5]For additional research on the relationship between physical motion and mental simulation, see DeLucia and Liddell (1998); Finke and Pinker (1982); Hubbard (1995, 2005).

[6]A fairly common experience at pool tables in London.



(a) A brick in front of the gate

(b) A puddle of beer in front of the gate

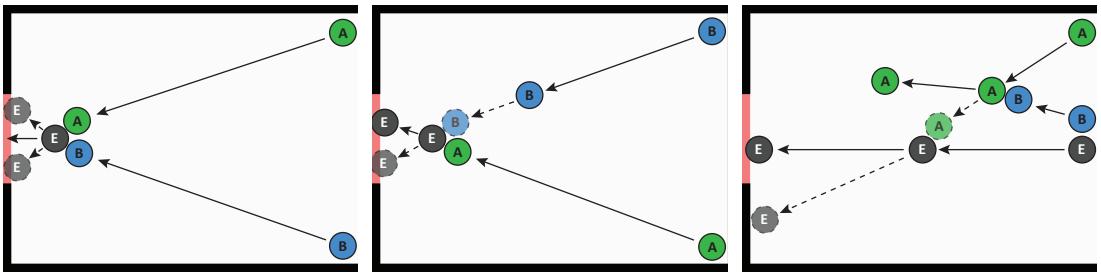
Figure 6.14: Two problem cases for Wolff's (2007) model.

forces is such that Wolff's account predicts that A helped (or enabled) B to reach the endstate. As mentioned above, in its current specification, the predictions of Wolff's account are insensitive to the magnitude of the different forces involved – only the configuration of forces matters. However, as we have seen above, when A only somewhat speeded up B (see clip 18 in Figure 6.6) most people judged that A neither caused nor prevented B (Experiment 3) or indicated that A had no significant effect on whether B went through the gate (Experiment 4). In this situation, B would have gone through the gate no matter whether or not A would have been present.

However, the situation is different with a puddle of beer in front of the gate. Note that the configuration of forces (and even the magnitude of forces) at the time of collision could indeed be identical to the situation without the puddle. Nevertheless, our account predicts that people's beliefs about how much the puddle will slow down B will affect their causal attribution. More precisely, we predict that when people are quite sure that B would *not* have passed the puddle without the additional speed gained through the collision with A, they will say that A *caused* B to go in. When people are unsure about whether B might have made it by itself, they will say A *helped* B and when they are sure that B would have gone in anyhow people are predicted to say that A *had no significant effect* on whether B went in. In order to derive these predictions from our model, we have to introduce noise not only in terms of *which direction* exactly a ball will go (or would have gone) but also in terms of *how fast* a ball will go (or would have gone, see Smith & Vul, 2012). Introducing noise on the magnitude of a ball's velocity would also allow our account to predict that A will be said to have *almost caused* B to go in the gate when B just comes to stop at the very end of the beer puddle.[7]

Having established the close fit of our model with people's judgments for relatively simple interactions, we will increase the complexity of the experimental domain in future work. With two cause balls, A and B, and one effect ball E (see Figure 6.15) we

[7] I am grateful to Chris Carroll for having raised this point.



(a) Overdetermination
A and B are each individually sufficient to cause E to go through the gate.

(b) Preemption
E is knocked in by A. B would have knocked in E a moment later.

(c) Double prevention
B prevents A from preventing E from going through the gate.

Figure 6.15: Complex interactions between physical objects.

can already model many of the cases that were traditionally deemed problematic for counterfactual theories of causal attribution. In the preemption case (Figure 6.15b), for example, the question is whether people's causal attributions to ball A are influenced by the fact that B would have also caused E to go through the gate somewhat later. While previous research has relied mostly on scenario-based paradigms (e.g. Walsh & Sloman, 2011), our physics world allows us to investigate these cases in a more controlled fashion.

6.2 Productive Concept Use (Gerstenberg & Goodman, 2012)

In the previous section, we have seen that people use their intuitive understanding of physics to make causal attributions. People's capacity to make judgments about what will happen or what might have happened (as well as the causal attributions which are informed by these judgments) go beyond the computational expressivity of static causal Bayes nets. In this section, we will see another way in which people's inferences go beyond what can be modelled with simple Bayes nets.

People often make surprisingly accurate inferences about a person's latent traits from very sparse evidence. If NG loses to TG in a ping pong match and afterwards wins against two other lab members, we are fairly confident that TG is a strong player despite only having observed him winning a single game. However, if we consequently find out that NG felt a bit lazy in his match against TG and did not try as hard as he normally does, our belief about TG's strength might change. This reasoning is not limited to a particular set of potential players, it can be generalised to related situations (such as team matches), and it supports inferences from complex combinations of evidence (e.g. learning that NG was lazy whenever he played a match against a team that included TG) – human reasoning is remarkably *productive*.

How can we best model the flexible inferences people draw from diverse patterns

of evidence such as the outcomes of matches in a ping pong tournament? What assumptions about the cognitive system do we need to make to be able to explain the productivity and gradedness of inference? What is the minimum level of abstraction that mental representations need to exhibit in order to support the inferential flexibility that our cognitive machinery displays?

There are two traditional, but fundamentally different ways of modelling higher-level cognition, each with its own strengths and drawbacks: statistical approaches (e.g. Rumelhart & McClelland, 1988) support graded probabilistic inference based on uncertain evidence but lack some of the representational powers of more richly structured symbolic approaches. Symbolic approaches (e.g. Newell, Shaw, & Simon, 1958), on the other hand, are confined to operating in the realm of certainty and are ill-suited to modelling people's inferences in a fundamentally uncertain world. More recently, researchers have started to break the dichotomy between statistical and symbolic models (Anderson, 1996) and have shown that much of cognition can be understood as probabilistic inference over richly structured representations (Tenenbaum et al., 2011).

For instance, causal Bayesian networks (CBN; Pearl, 2000) have been proposed as a modelling framework that combines the strengths of both statistical and symbolic approaches. Given a particular representation of a task that the cognitive system faces, a CBN supports inferences about the probability of competing hypotheses for many different patterns of evidence. However, a CBN is limited to the specific situation it was designed to model, allowing inferences from different observations of existing variables, but not from fundamentally different combinations of objects or events.[8] While some attempts have been made to model more abstract knowledge by constructing CBNs with richer, hierarchical structures (Kemp & Tenenbaum, 2009) or by combining CBNs with propositional logic (Goodman, Ullman, & Tenenbaum, 2011; Griffiths, 2005), CBNs have only coarse-grained compositionality insufficient to support productive extensions over different objects and situations.

Human thought, in contrast, is characterised by an enormous flexibility and productivity (Fodor, 1975, 1983). We can flexibly combine existing concepts to form new concepts and we can make use of these concepts to reason productively about an infinity of situations. The *probabilistic language of thought* (PLoT) hypothesis (Goodman & Tenenbaum, in prep) posits that mental representations have a language-like compositionality, and that the meaning of these representations is probabilistic, allowing them to be used for thinking and learning by probabilistic inference. This view of the representation of concepts provides a deeper marriage of the statistical and symbolic view. Because they are probabilistic, they support graded reasoning under uncertainty. Because they are language-like, they may be flexibly recombined to productively describe new situations. For instance, we have a set of concepts, such as 'strength' and

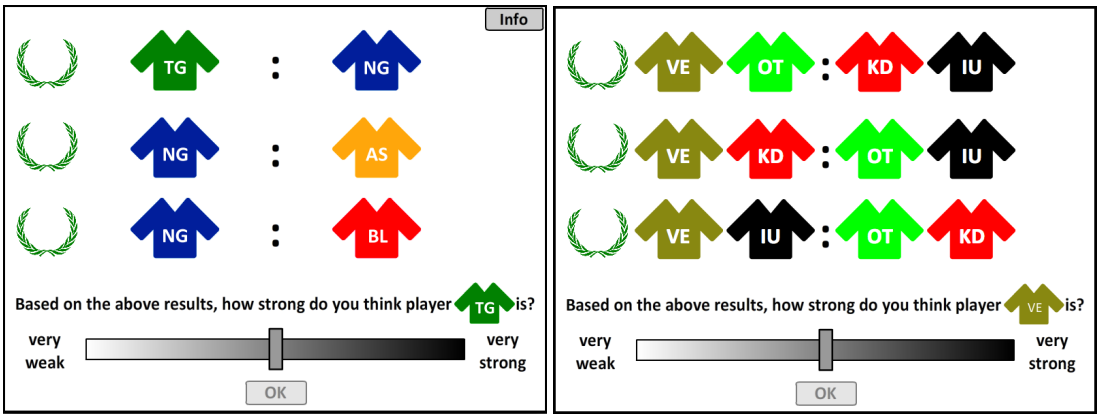
[8] Recall my criticism of how the CBN approach captures intuitions about who actually caused the bottle to shatter in the Billy and Suzy preemption scenario described at the beginning of this chapter.

'game', in the ping pong domain that we may compose together and apply to symbols such as TG. These combinations then describe distributions on possible world states, which we may reason about via the rules of probability. The PLoT hypothesis has been realised in existing computational systems, including the probabilistic programming language *Church* (Goodman et al., 2008). Church has several features that enable it to model productive inference from a small set of concepts – in particular, it allows reasoning about placeholder symbols and the forming of complex evidence by composing the concepts.

Here, we use Church (Goodman et al., 2008), as an instantiation of the PLoT, to explain aspects of people's flexible concept use, and use the ping pong scenario as a simple case study to illustrate our key points while admitting quantitative empirical evaluation. In two separate experiments, we test the predictions of our modelling approach by examining people's inferences based on complex patterns of causal evidence. We conclude by pointing out areas of research that are likely to benefit from this modelling framework.

6.2.1 Modelling probabilistic inferences in Church

Figure 6.16 shows two examples of the inference task that participants faced in the experiments which we will describe below. What representation would be needed to (a) be sensitive to the statistical nature of the evidence and (b) capture the abstract, symbolic structure that remains invariant between this particular situation and other similar situations that could involve different players and different outcomes? Figure 6.17 shows the Church code that we used to model people's inferences about a player's strength based on the results of ping pong tournaments. We chose the ping pong environment because it can be summarised by a relatively simple but rich set of concepts that support productive inferences from a variety of evidence in a variety of situations.



(a) Singles **(b)** Doubles

Figure 6.16: Screenshots of (a) single player tournament and (b) two player tournament. The winner of each match is indicated by a laurel wreath.

```
(mh-query 1000 100 ;Monte Carlo Inference
  ;CONCEPTS
  (define personstrength (mem (lambda (person) (gaussian 10 3))))
  (define lazy (mem (lambda (person game) (flip 0.1))))
  (define (teamstrength team game)
    (sum (map (lambda (person)
                (if (lazy person game)
                   (/ (personstrength person) 2)
                   (personstrength person)))
              team)))
  (define (winner team1 team2 game)
    (if (< (teamstrength team1 game)
        (teamstrength team2 game))
        'team2 'team1))
  ;QUERY
  (personstrength 'A)
  ;EVIDENCE
  (and
    (= 'team1 (winner '(TG) '(NG) 1))
    (= 'team1 (winner '(NG) '(AS) 2))
    (= 'team1 (winner '(NG) '(BL) 3))
    (lazy '(NG) 1) ;additional evidence, used in Experiment 2
  )
)
```

Figure 6.17: Church model of the ping pong scenario.

We will first introduce the Church language and then explain how this representation captures our intuitive concepts of ping pong.

Church is based on the λ-calculus, with a syntax inherited from the LISP family of languages (McCarthy, 1960). Thus operators precede their arguments, and are written inside grouping parentheses: `(+ 1 2)`. We use `define` to assign values to symbols in our program and `lambda` for creating functions. We could, for example, create a function `double` that takes one number as an input and returns its double. The code would look like this: `(define double (lambda (x) (+ x x)))`. What differentiates Church from an ordinary programming language is the inclusion of random primitives. For example, the function `(flip 0.5)` can be interpreted as a simple coin flip with a weight outputting either `true` or `false`. Every time the function is called, the coin is flipped afresh. A Church program specifies not a single computation, but a distribution over computations, or sampling process. This *sampling semantics* (see Goodman et al., 2008, for more details) means that composition of probabilities is achieved by ordinary composition of functions, and it means that we may specify probabilistic models using all the tools of representational abstraction in a modern programming language.

We now turn to describing the concepts (see `CONCEPTS` in Figure 6.17) that are required to represent the ping pong domain (Figure 6.16). This simple sports domain is built around people, teams and games. In Church, we can use symbols as placeholders for unspecified individuals of these types. This means that we do not need to define in advance how many people participate, what the size of the teams will be, or how many games a tournament will have. We define an individual player's strength, `personstrength`, via a function that draws from a Gaussian distribution with $M = 10$

and $SD = 3$. The memoization operator `mem` ensures that the strength value assigned to a person is persistent and does not change between games. We next make the assumption that players are sometimes `lazy`. The chance of a person being lazy in a particular game is 10%, specified by using the function `flip` with a weight of 0.1. As mentioned above, we also want to allow for the possibility that individual players form teams – we thus need the overall strength of a team, `teamstrength`. Here, we define the team's strength as the sum of the strength of each person in the team. If a person in the team is lazy, however, he only plays with half of his actual strength. Notice that this specification of the concept `teamstrength` is agnostic to the size of the team, and hence extends over a great variety of situations.

The way in which we can define new concepts (e.g. `teamstrength`) based on previously defined concepts (`personstrength` and `lazy`) illustrates the compositionality of Church. Finally, we specify how the `winner` of a game is determined. We simply say the team wins which has the greater overall strength. This set of function definitions specifies a simple lexicon of concepts for reasoning about the ping pong domain. The functions are built up compositionally, and may be further composed for specific situations (see below). Moreover, the set of concept definitions refers to people (teams, etc.) without having to declare a set of possible people in advance: instead we apply generic functions to placeholder symbols that will stand for these people. Table 6.2 concisely summarises our modelling assumptions.

Now we have a lexicon of concepts (`CONCEPTS`) that we may use to model people's inferences about a player's strength (`QUERY`) not only in the situations depicted in Figure 6.16 but in a multitude of possible situations with varying teams composed of several people, playing against each other with all thinkable combinations of game results in different tournament formats (`EVIDENCE`). This productive extension over different possible situations including different persons, different teams and different winners of each game, renders the Church implementation a powerful model for human reasoning.

A program in Church can be seen as a formal description of the process that gener-

Table 6.2: Modelling assumptions.

concept	description	assumption
`personstrength`	strength of a player	normally distributed, persistent property
`lazy`	chance that a player is lazy	$p(\text{lazy}) = 10\%$, not persistent
`teamstrength`	strength of a team	individual strengths combine additively
`winner`	winner of a match	team with greater strength wins

ates observed or hypothesised evidence. The `mh-query` operator specifies a conditional inference. Both the evidence provided and the question we are asking are composed out of the concepts that specify the domain. Church completely separates the actual process of inference from the underlying representations and the inferences they license.

This allows the modeller to focus on defining the conceptual representation of the domain of interest without having to worry about the exact details of how inference is carried out; it also provides a framework for psychological investigation of representations and the inferences that may be drawn, without committing to *how* these inferences are made – a well-formed level of analysis between Marr's computational and algorithmic levels (Marr, 1982). Hence, in contrast to other frameworks for building psychological models of cognition, such as ACT-R (Anderson, 1996), Church does not incorporate any assumptions about how exactly the cognitive system carries out its computations but postulates that inference accords with the rules of probability (see Griffiths, Vul, & Sanborn, 2012, for some ideas about how the computational and algorithmic level may be linked).

6.2.2 Experiment 1: Bayesian ping pong

In Experiment 1, we wanted to explore how well our simple Church model predicts the inferences people make, based on complex patterns of evidence in different situations. Participants' task was to estimate an individual player's strength based on the outcomes of different games in a ping pong tournament. Participants were told that they will make judgments after having seen single player and two-player tournaments. The different players in a tournament could be identified by the colour of their jersey as well as their initials. In each tournament, there was a new set of players. Participants were given some basic information about the strength of the players which described some of the modelling assumptions we made (see Table 6.2). That is, participants were told that individual players have a fixed strength which does not vary between games and that all of the players have a 10% chance of not playing as strongly as they can in each game. This means that even if a player is strong, he can sometimes lose against a weaker player.

6.2.2.1 Method

Participants 30 (22 female) recruited through Amazon Mechanical Turk participated in the experiment. The mean age was 31.3 years ($SD = 10.8$).

Materials and Procedure The experiment was programmed in Adobe Flash CS5.[9] Participants viewed 20 tournaments in total. First, one block of 8 single player tournaments and then another block of 12 two-player tournaments. The order of the tourna-

[9]Demos of both Experiments can be accessed here:
`http://www.ucl.ac.uk/lagnado-lab/experiments/demos/BPP_demos.html`

Table 6.3: Patterns of observation for the single player tournaments. *Note*: An additional set of 4 patterns was included for which the outcomes of the games were reversed. The bottom row shows the omniscient commentator's information in Experiment 2.

confounded evidence (1,2)	strong indirect evidence (3,4)	weak indirect evidence (5,6)	diverse evidence (7,8)
A > B	A > B	A > B	A > B
A > B	B > C	B < C	A > C
A > B	B > D	B < D	A > D
lazy,game: B,2	B,1	B,1	C,2

Note: A > B means that A won against B.

ments within each block was randomised. Participants could remind themselves about the most important aspects of the experiment by moving the mouse over the 'Info' field on the top right of the screen (see Figure 6.16). Based on the results of the three matches in the tournament, participants estimated the strength of the indicated player on a slider that ranged from -50 to 50. The endpoints were labelled 'very weak' and 'very strong'. On average, it took participants 7.4 ($SD = 3.3$) minutes to complete the experiment.

Design Table 6.3 shows the patterns of evidence that were used for the single player tournaments. Table 6.4 shows the patterns for the two-player tournaments. In all tournaments, participants were asked to judge the strength of player A.

For the single player tournaments, we used four different patterns of evidence: *confounded evidence* in which A wins repeatedly against B, *strong* and *weak indirect evi-*

Table 6.4: Patterns of observation for the two-player tournaments. *Note*: An additional set of 6 patterns was included in which the outcomes of the games were reversed.

confounded with partner (9,10)			confounded with opponent (11,12)			strong indirect evidence (13,14)		
AB	>	CD	AB	>	EF	AB	>	EF
AB	>	EF	AC	>	EG	BC	<	EF
AB	>	GH	AD	>	EH	BD	<	EF
weak indirect evidence (15,16)			**diverse evidence (17,18)**			**round robin (19,20)**		
AB	>	EF	AB	>	EF	AB	>	CD
BC	>	EF	AC	>	GH	AC	>	BD
BD	>	EF	AD	>	IJ	AD	>	BC

Note: AB > CD means that the team with players A and B won against the team with players C and D.

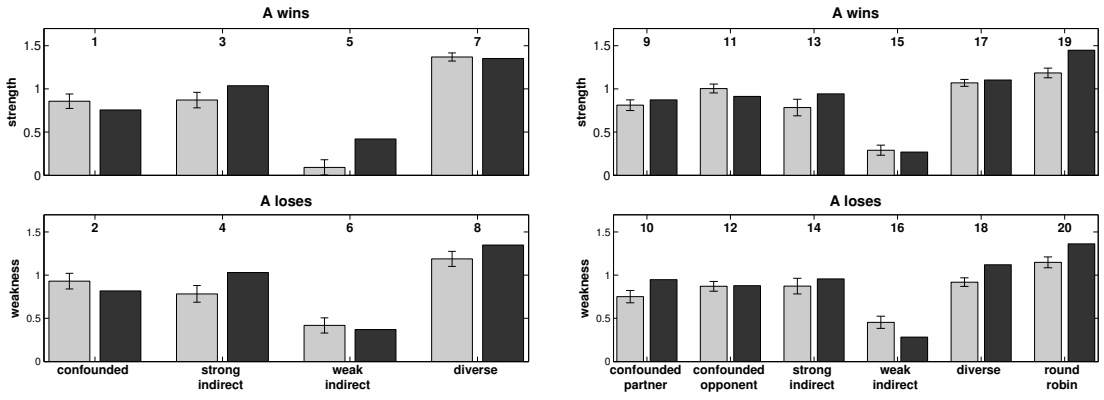
dence where A only wins one match herself but B either continues to win or lose two games against other players and *diverse evidence* in which A wins against three different players. For each of those patterns, we also included a pattern in which the outcomes of the games were exactly reversed.

For the two player tournaments, we used six different patterns of evidence: In some situations A was always in the same team as B (*confounded with partner*) while in other situations A repeatedly played against the same player E (*confounded with opponent*). As in the single player tournaments, we also had patterns with mostly indirect evidence about the strength of A by having his partner in the first game, B, either win or lose against the same opponents with different teammates (*weak/strong indirect evidence*). Finally, we had one pattern of *diverse evidence* in which A wins with different teammates against a new set of opponents in each game and one *round robin* tournament in which A wins all his games in all possible combinations of a 4-player tournament.

6.2.2.2 Results and discussion

In order to directly compare the model predictions with participants' judgments we z-scored the model predictions and each individual participant's judgments. Furthermore, we reverse coded participants' judgments and the model predictions for the situations in which the outcomes of the games were reversed so that both strength and 'weakness' judgments go in the same direction.

Figure 6.18 shows the mean strength estimates (grey bars) together with the model predictions (black bars) for the single and two-player tournaments. The top panels display the situations in which A won his game(s). The bottom panels show the situations in which A lost. Our model predicts participants' judgments in the single and two-player tournaments very well with $r = .98$ and $RMSE = .19$. A very high median correlation



(a) Single player tournament **(b)** Two-player tournament

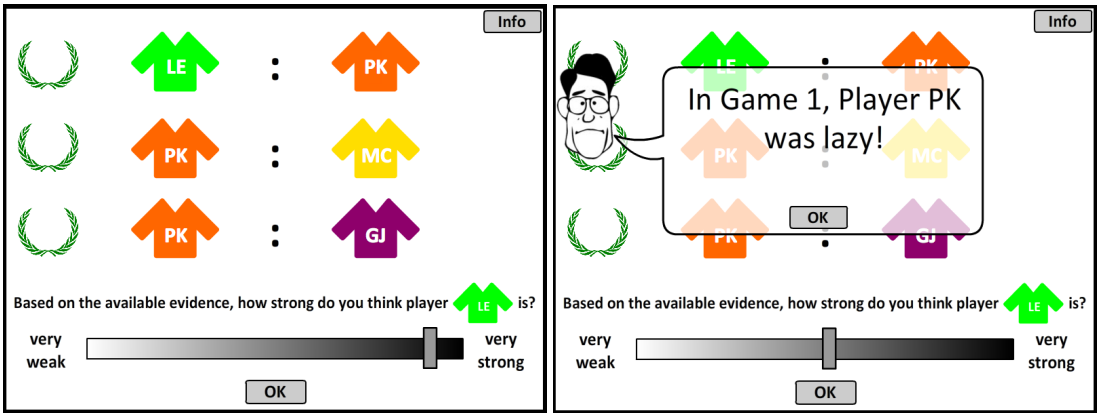
Figure 6.18: Z-scored mean strength estimates (grey bars) and model predictions (black bars) for the single player (left) and two-player tournaments (right). Numbers above the bars correspond to the patterns described in Tables 6.3 and 6.4. Error bars are ± 1 *SEM*.

with individual participants' judgments of $r = .92$ shows that the close fit is not merely due to an aggregation effect.

In describing the data qualitatively, we will focus on the strength jugdments in the top panels (strength and weakness judgments were highly correlated, $r = .96$). In the single player tournaments, A is judged equally strong when he repeatedly wins against the same player (Situation 1) or when strong indirect evidence was provided (3). A is judged weakest when only weak indirect evidence is provided (5). A is judged to be strongest when she won against three different players (7). In the two-player tournaments, A is judged equally strong when the evidence is confounded with the partner or opponent and when strong indirect evidence is provided (9, 11 and 13). A is judged to be relatively weak when only weak indirect evidence is provided (15). A is judged to be strong for the situations in which participant's received diverse evidence about A's strength (17) and even stronger for the round robin tournament (19).

There appears to be only one prediction that the model makes which is not supported by the data. In the single player tournaments, the model predicts that participants should be slightly more confident about the strength of A when provided with strong indirect evidence (Situations 3, 4) compared to when confounded evidence is given (Situations 1, 2). However, there is no significant difference between participants' judgments for strong indirect evidence ($M = 26.2$, $SD = 15.4$) compared to confounded evidence ($M = 27.8$, $SD = 13.8$), $t(29) = 0.44, p > .05$.

The results of Experiment 1 show that our model predicts participants' inferences very accurately. We have demonstrated that a single and concise representation of the task is sufficient to predict people's inferences for a great diversity of patterns of evidence. If we were to model participants' inferences with the use of simple CBNs, we would have needed to construct a separate CBN for each tournament. The close fit between our model and participants' inference also shows that our modelling assumptions (e.g. that the team's strength is a linear combination of the individual team members' strengths) generally matched participants' implicit assumptions (see Table 6.2). However, the fact that the model's prediction of a difference between strength judgments based on strong indirect evidence versus confounded evidence was not supported by the data, suggests that participants might have differed in the extent to which they took the chance of laziness into consideration. In fact, only 16 out of 30 participants showed the pattern in the predicted direction. If we increase the probability of a person being lazy in a particular game in the model, it matches participants' average judgments for these situations. Intuitively, if the chances of a person having been lazy in a particular game are increased, there is a higher chance that player A won his game against player B in Situation 3 because B was lazy in this round. However, when A wins repeatedly against B, there is hardly any effect of changing the probability of laziness. For example, it is very unlikely when A won three times against B, that B (and not A) was lazy three times in a row.



(a) Match information only **(b)** Omniscient commentator

Figure 6.19: Screenshots of the omniscient commentator condition.

6.2.3 Experiment 2: Omniscient commentator

In Experiment 1 we have shown that our model accurately predicts participants' inferences for a great variety of patterns of evidence from different combinations of teams and outcomes. A still greater variety of evidence is available by composing the basic concepts together in different ways: there is no reason for evidence not to directly refer to a player's strength, laziness, etc. While in Experiment 1, the match results were the only source of information participants could use as a basis for their strength judgments, Experiment 2 introduced an omniscient commentator who gave direct information about specific players. After participants saw a tournament's match results, an omniscient commentator, who always told the truth, revealed that one player was lazy in a particular game (see Figure 6.19).

We were interested in how participants updated their beliefs about the strength of player A given this additional piece of evidence. Importantly, we do not need to change anything in the Church code to derive predictions for these situations since all the necessary concepts are already defined.

6.2.3.1 Method

Participants 20 (11 female) participants recruited through Amazon Mechanical Turk participated in the experiment. The mean age was 34 years ($SD = 9.8$).

Materials, Procedure and Design Participants viewed 10 single player tournaments which comprised the 8 situations used in Experiment 1 plus two additional patterns (IR 1, 2). Participants first judged player A's strength based merely on the match results in the tournament (see Figure 6.19a). Afterwards, participants received information from the omniscient commentator about one player who was lazy in a particular match 6.19b. Participants then rated A's strength for a second time, whereby the slider

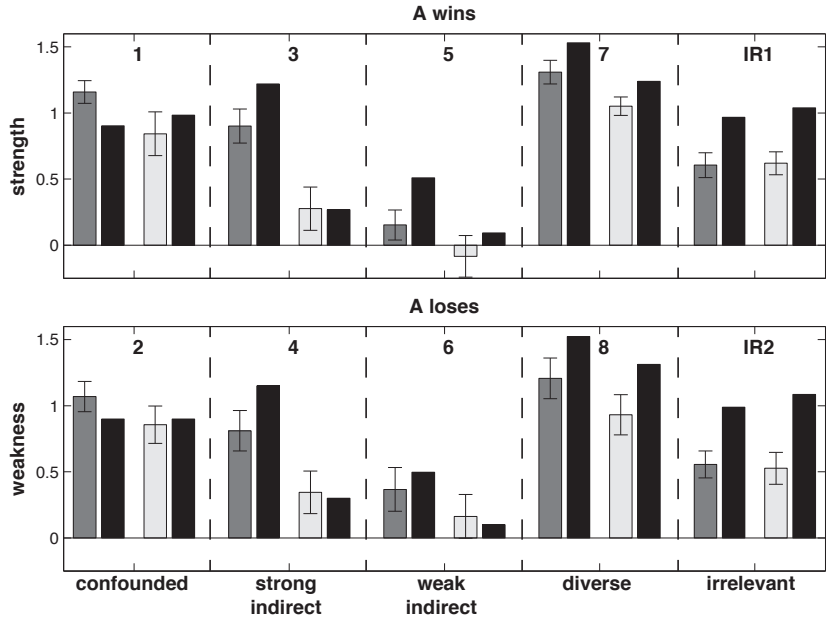


Figure 6.20: Z-scored mean strength estimates and model predictions. Dark grey bars = estimates after tournament information only, light grey bars = estimates after omniscient commentator info, black bars = model predictions. Error bars are ± 1 *SEM*.

was initialised at the first judgment's position. On average, it took participants 9.4 ($SD = 4$) minutes to complete the experiment.

The bottom row of Table 6.3 shows what information the omniscient commentator revealed in each situation. For example, in situation 3 in which participants first saw strong indirect evidence, the commentator then said: "In game 1, Player B was lazy." In the additional pattern (IR 2), A wins against B, B wins against C and D wins against E. The commentator then reveals that E was lazy in game 3. For the patterns in which A lost his game, the results of each match as shown in Table 6.3 were reversed and the corresponding losing player was indicated as having been lazy. For example, in situation 2, A lost all three games against B and the commentator revealed that A was lazy in game 2.

6.2.3.2 Results and discussion

Figure 6.20 shows the mean strength judgments (grey bars) together with the model predictions (black bars). The dark grey bars indicate participants' first judgments based on the tournament information only. The light grey bars indicate participant's second judgments after they received the commentator's information. The model predicts participants' ratings very accurately again with $r = .97$ and $RMSE = 0.29$. The model's median correlation with individual participants' judgments is $r = .86$. Again, strength and weakness judgments for the corresponding patterns were highly correlated, $r = .98$.

Generally, participants lowered their estimate of A's strength (top panel) and weak-

ness (bottom panel) after having received the commentator's information. The fact that participants do not lower their estimates of A's strength for the two cases in which they received *irrelevant evidence* by the commentator about a player' laziness who was in no relationship with A (IR 1, 2), shows that participants did not just have a tendency to regress towards the mean of the scale in their second judgments.

As predicted by the model, the degree to which participants lowered their strength estimates as a result of the laziness information differed between situations. While participants only marginally lowered their estimates for the *confounded evidence* patterns, estimates went down considerably for the *strong indirect evidence* patterns. As mentioned in the discussion of Experiment 1, finding out in the *strong indirect evidence* situation that A's win against B might have only been due to the fact that B was lazy in this match undermines the relevance of the additional evidence about B's performance in match 2 and 3 for A's strength.

The results of Experiment 2 show that participants, as well as our model, have no difficulty in integrating different sources of evidence to form an overall judgment of a player's likely underlying strength. The model predicts participants' judgments very accurately by being sensitive to the degree to which the initial strength estimate should be updated in the light of new evidence provided by the commentator.

6.2.4 General discussion

In this section, we have demonstrated a novel modelling framework that conceptualises people's reasoning as probabilistic inference over compositionally structured representations. With a handful of concepts that can combine compositionally and support productive extensions over novel situations and objects, we predict participants' judgments in two experiments with thirty different patterns of evidence in total.

The fact that people can reason flexibly based on different patterns and sources of evidence illustrates the importance of modelling our representational capacities on a sufficiently abstract level. People's use of concepts is not tied to particular situations but extend productively over different contexts. The concept of a *winner*, for example, applies to a whole range of possible games or even to domains outside of games entirely such as winning an election. We have provided a concrete working-example of how such a representation could look like, using the probabilistic programming language Church (Goodman et al., 2008). The fact that our model's predictions corresponded very closely to people's judgments can be taken as evidence that the assumptions we had to make when writing the program, generally matched the intuitive assumptions that people brought to the task. A Church program makes the modelling assumptions explicit and thus allows them to be scrutinised. Furthermore, particular modelling assumptions can also be treated as parameters in the model. For example, as outlined above, different participants seemed to have given unequal weight to the probability that a player might

be lazy in a game. Without changing the general structure of our representation, we could account for these individual differences by allowing for flexibility in our modelling assumptions through treating the chance of laziness as a free parameter.

In our experiments, we have focused on a single query and only used a small number of the possible patterns of evidence. However, our representation supports many more combinations of queries and evidence. For example, we could ask about the probability that a particular player was lazy in a certain game. Or we could ask which of two teams is likely to win given that we have observed the players perform in some previous games or based on some direct information about their strength. Furthermore, it would require only minimal additions to the concept lexicon to handle evidence such as, "all players in the red jerseys were lazy" or "at least one of the players in the green jerseys is very strong."

To conclude, we have provided only a small glimpse into what we see as a broad research program that investigates people's flexible use of everyday concepts using the tools of probabilistic programming – the probabilistic language of thought hypothesis. We believe that this research program has the potential to greatly benefit our understanding of how higher-level capacities of human cognition (such as concept learning, naive physics, and theory of mind) are possible (see, e.g. Goodman & Stuhlmüller, 2012).

6.3 Conclusion

In this chapter, we have seen two ways in which human cognition goes beyond what can be expressed with simple Bayes nets. First, people possess an intuitive understanding of physics and use this knowledge to simulate possible outcomes which inform their causal attributions. Second, people's inferential abilities exhibit a flexibility and generality that cannot be captured within a CBN framework. We have shown that a modelling approach based on the assumption that people's mental representations are compositional and probabilistic predicts people's inferences about a latent continuous variable based on complex patterns of evidence very accurately.

I believe that both strands of research bear great potential for further exploration. In particular, it will be interesting to see how both strands can be combined to derive predictions about people's dispositional inferences and causal attributions when both physical and social evidence need to be combined. Understanding causal attributions in terms of counterfactuals defined over probabilistic, generative models (which can be precisely formulated as probabilistic programs) combines what we see as the positive features of dependency and process accounts of causation. Our framework exhibits the generality of the dependency accounts and can thus be applied to situations in which our understanding of the underlying processes is limited. Of course, the less knowledge we have of how different variables in the system are interconnected, the

less likely we are to make veridical attributions. For example, a person who has only a limited understanding of how economical markets or political systems work, is less likely to identify the correct sources of problems and suggest appropriate levers for intervention. On the other hand, the more certain we are about how the underlying causal system works, the more confident we should be in our causal attributions.

Thanks to modern software packages which implement physical laws, we now have the resources to design sophisticated experiments and evaluate participants' attributions in a quantitative fashion based on simulations within this environment. Whereas most research in attribution up until now has almost exclusively relied on a scenario-based approach, these novel tools enable us to generate much more realistic stimuli. In the real world, people rarely make attributions based on descriptions only but rather on direct observation of what happened in a particular situation. Even in courts, sophisticated computer simulations that go beyond mere descriptions of the crime are used more and more frequently.[10]

I consider the prospect of investigating how people combine physical and social cues to make causal and responsibility attributions to be particularly exciting (see Iliev, Sachdeva, & Medin, 2012; Zhou et al., in press). Einhorn and Hogarth (1986, p. 3) have argued that "discussions of how people perceive physical causation are rarely linked to the processes of causal attribution in the social domain, or vice versa." In a recent experiment, Iliev et al. (2012) used video clips in which two geometrical objects interacted with each other. In all clips, there was an actor and a victim which was harmed. Participants were asked to judge for pairs of clips in which out of the two clips the actions of the actor were worse. In their clips, they varied several factors such as the distance the actor travelled to harm the victim, the force with which the two interacted, the duration (and frequency) of contact between actor and victim, and whether the actor intervened on the victim directly or on the source of harm. Overall, participants judged interventions on the victim to be worse than interventions on the source of harm (cf. Waldmann & Dieterich, 2007). Furthermore, setting an object in motion was seen as worse compared to redirecting an object that already is in motion. Finally, participants were also sensitive to the duration and frequency of interactions with more and longer interactions being seen as morally worse than shorter or fewer interactions.

The clips that Iliev et al. used left little uncertainty about what would have happened in the absence of the actor's actions. It would be interesting to apply our model of causal attribution from Section 6.1 to this setup. Our model would predict that people's beliefs about what would have happened in the absence of the actor's action, will modulate people's perception of how bad the action was. For example, if participants are sure that the victim would not have suffered but for the actor's action they should evaluate the

[10]For a reconstruction of the assassination of John F. Kennedy see `http://www.jfkfiles.com/` (retrieved on 14/12/2012) and for a short clip demonstrating the simulation see `http://www.youtube.com/watch?v=DSBXW1-VGmM` (retrieved on 14/12/2012).

actor maximally negatively. However, when they think that the same negative outcome would have prevailed even without the intervention of the actor, then the actor should be judged less harshly. This might explain, for example, why setting in motion was perceived as worse compared to the redirection of movement. A moving agent is more likely to fall into a trap (especially when moving vaguely in that direction) compared to an agent who is resting. Furthermore, it would be particularly interesting to explore what happens when multiple actors interact with each other to bring about a negative outcome. Future research could go even one step further by leaving people initially uncertain about the underlying intentions of different agents. Similar to the classic Heider and Simmel (1944) stimuli, participants would learn about relevant dispositional states (such as an agent's intention and their ability/strength) through observing agents interacting in one situation and then make causal (or moral) attributions in another situation.

People have rich intuitive theories about how the physical and social world in which they are embedded works (e.g. Griffiths & Tenenbaum, 2009; Tenenbaum et al., 2006, 2007, 2011). In this chapter, we have seen two examples for why it will be necessary to move beyond simple Bayesian networks in order to adequately model people's attributions and inferences. I am excited about the prospects of merging what we have learned from social psychology with sophisticated modelling techniques in the cognitive sciences to develop formal models of social cognition – models that acknowledge the diversity of evidence (both physical and social) on which people's attributions are based.

Chapter 7

Summary and Conclusions

> "Man is so intelligent that he feels impelled to invent theories to account for what happens in the world. Unfortunately, he is not quite intelligent enough, in most cases, to find correct explanations. So that when he acts on his theories, he behaves very often like a lunatic."
>
> – **Aldous Huxley**

In this thesis, I have developed a novel framework of responsibility attribution. In this framework, attributions of responsibility are modelled in terms of counterfactuals defined over a causal representation of the situation. This way of conceptualising responsibility combines insights from formal models in social psychology (Brewer, 1977; Fincham & Jaspars, 1983; Petrocelli et al., 2011; Spellman, 1997) with recent developments in computer science (Chockler & Halpern, 2004; Pearl, 2000; Spirtes et al., 2000) and philosophy (Hitchcock, 2001b; Woodward, 2003; Yablo, 2002). The main intuition is that there is a close relationship between the extent to which a person's action influenced the outcome and their degree of responsibility. In our framework, we model a person's influence via comparing what actually happened with what would have happened if the person's action had been different. All else being equal, responsibility increases the more a person's action was perceived to have made a difference to the outcome. We assume that people employ their causal understanding of the situation to reason not only about what actually happened but also about what would have happened under different counterfactual contingencies.

Responsibility cannot simply be equated with counterfactual dependence: people are sometimes held responsible even if their action made no difference to the outcome. In order to accommodate this finding, our framework relies on a relaxed notion of counterfactual dependence (cf. Chockler & Halpern, 2004; Halpern & Pearl, 2005; Woodward, 2003). Even if a person's action made no difference in the actual situation, she can still be held responsible for the outcome *if* it was possible for a situation to occur in which her action would have made a difference. Our framework predicts that responsibility

decreases the more events in the actual world would have needed to change in order to generate a counterfactual world in which the person's action would have made a difference. The predictions of this framework were subsequently tested in a series of empirical studies.

7.1 Summary of the Main Findings

In Chapter 4, I have discussed three sets of studies in which we investigated how participants' responsibility attributions to individuals within a group are influenced by the underlying causal structure of a situation. The results of all three studies showed that the nature of the causal structure has a marked effect on how responsibility is distributed.

In Section 4.1, we demonstrated how responsibility attributions are influenced by exactly how the individual contributions combine to determine the group outcome (Gerstenberg & Lagnado, 2010). We manipulated whether the performance of each individual influenced the group outcome in an additive, conjunctive or disjunctive manner. The results showed that participants' responsibility attributions were not only affected by performance but moderated by the causal structure. In line with the predictions of our framework, individuals received less responsibility in situations in which the outcome was overdetermined. In contrast to the predictions of a simple counterfactual model, responsibility did not reduce to zero in such situations. Rather, individual responsibility remained substantial but reduced the more the outcome was overdetermined. Overall, we found that the structural model (Chockler & Halpern, 2004) predicted participants' responsibility attributions better than a simple counterfactual model.

In Section 4.2, we looked at how responsibility diffuses between the members of a group in asymmetric causal structures (Zultan et al., 2012). In such structures, different group members' contributions influence the collective outcome to varying degrees. The results of these experiments provided further evidence for the importance of causal and counterfactual considerations for attributions of responsibility. In general, responsibility did not just diffuse equally between the group members. As predicted by the structural model, a person's responsibility was not only determined by their own performance but also by the performance of their group members *and* the relationships between the members. For example, if the group outcome was negative, responsibility attributions to one player tended to increase with the success of complementary players and decrease with the success of substitutes.

These effects are predicted by our general framework. With every successful substitute, the world in which the person of interest would make a difference moves further away from the actual world. In contrast, with every successful complement, the two worlds move closer together. However, the results also showed a systematic inconsistency with the predictions of Chockler and Halpern's (2004) structural model. Players who were equal in terms of pivotality (i.e. equally distant from a world in which their

contribution would have made a difference) received different degrees of responsibility. Part of this inconsistency was resolved by extending the structural model. According to the multiple-path model, responsibility is not only a function of how distant the closest possible world is but also by the number of possible worlds in which the person's contribution would have been pivotal. Another experiment supported the superiority of the multiple-path model over the structural model.

In Section 4.3, we subjected our framework to an even more rigorous test (Lagnado et al., accepted). Participants made responsibility attributions in a large variety of situations that manipulated the group structure and the patterns of performance. Both the structural model and the multiple-path model predict that a pivotal person will always be held fully responsible for the outcome. However, in contrast to this prediction, participants' responsibility attributions differentiated between pivotal players. Furthermore, players who were less pivotal were sometimes attributed *more* responsibility than pivotal players. These results demonstrate that participants' attributions are not exclusively determined by pivotality reasoning.

We developed a theoretical extension according to which responsibility attributions are also influenced by how critical a person's contribution was perceived for the positive outcome. All else being equal, responsibility is predicted to increase with criticality. The predictions of this general criticality-pivotality framework were borne out by the data. If criticality was held constant, responsibility increased with pivotality as predicted by both the structural and multiple-path model. However, in contrast to the structural and multiple-path model, responsibility attributions also increased with criticality when pivotality was held constant.

This result fits well with the general message that responsibility attributions are not solely determined by what actually happened. In fact, there are at least two ways in which counterfactual considerations affect attributions of responsibility – via pivotality *and* criticality. Given what actually happened, responsibility attributions are sensitive to how close a person was to making a difference. In addition to that, people are also concerned with the number of possible worlds in which a person's contribution would be critical for a positive group outcome.

In sum, the empirical findings in Chapter 4 support the validity of the framework. Attributions of responsibility can be modelled in terms of counterfactuals defined over a causal representation of the situation. Conceptualising responsibility in such a way highlights the intimate relationship between responsibility, causality and counterfactuals. People's attributions of responsibility are sensitive to what actually happened and to what could have happened in other possible worlds as dictated by the causal structure of the situation.

In Chapter 5, I demonstrated via a series of experiments how responsibility attributions are influenced by differences in mental states (Gerstenberg et al., 2011; Gerstenberg & Lagnado, 2012; Gerstenberg et al., 2010; Schächtele et al., 2011). Focussing

again on situations in which participants attribute responsibility to group members for collective outcomes, I showed how attributions were sensitive to differences in knowledge (Section 5.1), expectations (Section 5.2) and intentions (Section 5.3). Differences in mental states affect both pivotality and criticality – the central pillars of our general responsibility attribution framework.

In Section 5.1, we investigated how people attribute responsibility when the individual performances of players combine sequentially (Gerstenberg & Lagnado, 2012). The team succeeded in the challenge if the summed team members' scores exceeded a certain threshold. We manipulated whether or not the group outcome was already certain prior to the last player in the team. When the final player knew that the team's loss was already certain (because the threshold could not be reached anymore), blame attributions to the last player were reduced compared to situations in which the outcome was still undecided. Similarly, the last player received less credit when the team's win was already certain prior to her turn.

However, in contrast to what would be predicted from a model based on a simple counterfactual criterion (cf. Spellman, 1997), responsibility attributions to the last player for overdetermined outcomes were still sensitive to the player's performance. She received more blame and less credit for a poor performance compared to a good performance. We argue that this effect of performance on responsibility attributions in situations of overdetermination is in line with pivotality reasoning. Although the player's level of performance made no difference in the actual situation it could have made a difference in another possible situation in which the teammates' performances had been different.

The results also supported the importance of criticality considerations for responsibility attribution. In another experiment, we exactly replicated the structure and performances of the players. However, this time, participants were instructed that later players did not know about their earlier teammates' scores. The results showed that responsibility attributions to the final player were no more influenced by the performances of the previous players. Attributions to the final player were only reduced when that player was aware of the fact that her performance was not critical anymore for the team outcome.

In Section 5.2, we manipulated another factor that is predicted to have an influence on pivotality and criticality – the skill level of different players in a team (Gerstenberg et al., 2011). According to our general framework, varying one person's level of skill should influence another player's criticality. For example, the more skilled one partner is in a disjunctive challenge, the less critical the other partner should be seen for the team's positive outcome. For pivotality, our framework predicts that when two players failed, it should be easier to imagine that a player with a high skill level would have succeeded than a player with low skill.

In the experiment, we varied the players' skill level and performance. Naturally,

responsibility attributions were strongly influenced by performance. Players received more credit and less blame for better performances. Furthermore, attributions were also affected by the players' skill. While skilled players received more blame than unskilled players, both players received equal credit for positive outcomes.

This set of findings only provides partial support for our general framework. The fact that skilled players are blamed more than unskilled players is predicted by our framework. Because the nature of the task was disjunctive, criticality is predicted to increase with the relative difference in skill. However, our general framework cannot predict why there were no differences as a function of skill for positive outcomes. Additionally, the results failed to provide evidence for pivotality reasoning: participants' responsibility attributions were not affected by whether or not the outcome was overdetermined.

Some of the limitations that might have led to this mixed pattern of results are that first, participants were instructed that their attributions affect each player individually and second, participants were left unsure about whether the players in the team knew about their partner's skill level. The first aspect interferes with pivotality and the second with criticality. Investigating the effects of different priors on responsibility attributions remains an important avenue for future research.

In Section 5.3, the focus was on the effect of intentions on attributions of responsibility (Gerstenberg et al., 2010) and on interactions in an economic game (Schächtele et al., 2011). Two separate studies shared the general setup of employing a paradigm with a probabilistic relationship between intentions and outcomes. While the valence of intention and outcome is matched most of the time, a positive intention can sometimes lead to a negative outcome and vice versa. In both sets of experiments, we were interested to what extent participants' attributions were affected by the intended versus actual outcome.

In Section 5.3.1, participants evaluated the behaviour of different players in a group game with a classic social dilemma structure (Gerstenberg et al., 2010). If each person behaves according to what's in their best self-interest, the outcome for the group is likely to be worse than what it would be if individuals refrained from trying to maximise their individual gains. The results showed that participants took both the player's intended as well as their actual contributions into account when attributing responsibility. When looking at attributions on the level of individual participants, we found that the attributions of most participants were better explained by an intention-based rather than an outcome-based model. Indeed, there was a significant number of participants who only attributed responsibility as a function of the players's intentions and irrespective of the outcome.

In Section 5.3.2, we employed the same basic setup of a noisy relationship between intentions and outcomes (Schächtele et al., 2011). However, this time we looked into how economic interactions in dyads are affected by intentions and outcomes. Using an allocator-responder game, we found that responders' were affected by both the alloca-

tor's intended as well as the actual outcome. Similar to what we found in the previous section, responders were more strongly influenced by intentions than by outcomes. Moreover, responders behaviour was asymmetric: negative intentions were punished but positive intentions hardly rewarded.

In another experiment, we manipulated the responder's access to the allocator's choice. Whereas in the previous experiment the allocator's intention was observed, this time responder's only saw the outcome. The allocator then stated their intention. Thus, just like in real life, intentions were not directly observable but had to be inferred from the outcome together with the allocator's statement which could be deceptive. In fact, most allocators took advantage of the possibility to deceive about their real intentions. While the stated intentions were more generous than in the experiment in which intentions were observed, actual intentions were much more selfish. However, responders tended to mistrust allocators' statements. Again, (stated) intentions influenced responders' behaviour more strongly than outcomes. Even more interestingly, the effect of stated intentions on responder's behaviour was sometimes reversed. On experience of a negative outcome, responders punished allocators less who stated that the outcome resulted from a negative intention compared to allocators who claimed that their intention had been positive. Thus, there was a premium for honesty and perceived attempts of deception were punished harshly.

The effects of intentions again demonstrate the importance of counterfactual considerations for attributions of responsibility and the behaviour in economic games. Participants are not only concerned with the actual outcome but also with the probability of other outcomes given the same intention. In Section 5.4, I discussed how our general framework can be extended to incorporate situations in which different persons have different characteristics. As a first approximation, differences in intentions and skill level can be captured via assuming that these factors influence the prior chances that a certain outcome will come about. I have shown that a model of responsibility that uses Pearl's (2000) general framework for handling counterfactuals cannot predict effects of priors on responsibility attributions when there is no uncertainty over what values variables had in the actual world. However, a more flexible account which allows for a chance that events in the considered counterfactual worlds be different from what they actually were, yields precise predictions about how differences in priors should affect attributions of responsibility.

Finally, in Chapter 6, I have shown two domains in which people's inferential and attributional capabilities go beyond what can be modelled in terms of simple causal networks.

In Section 6.1, people's causal attributions for simple collision events were shown to be informed by their intuitive understanding of physics. We first provided evidence that people's predictions about possible future states are well accounted for by assuming that our intuitive understanding of physics in this domain is approximately Newtonian. In

line with the general framework developed in this thesis, we modelled participants' causal attributions in terms of counterfactuals defined over a causal model of the situation. People's causal attributions are a function of what actually happened and what they think would have happened in the absence of the causal event of interest. More precisely, causal attributions are closely related to people's subjective degree of belief over what would have happened in the relevant counterfactual world. An event A is seen more causal, the more certain a person is that another event B would not have happened but for A. Moreover, we showed how this epistemic approach to modelling people's causal attributions can be extended to yield quantitative predictions about when people will say that A helped rather than caused B or when A almost caused/prevented B. Together, these results highlight the generality of our general framework: it applies as long as we are of expressing people's domain knowledge in terms of probabilistic, generative models that support the simulation of counterfactuals.

Section 6.2 illustrated some of the advantages of modelling cognition in terms of probabilistic generative models. According to the probabilistic language of thought hypothesis (PLoT), thinking can be understood as probabilistic inference over compositionally structured representations. Using ping pong tournaments as a case study, we showed how a single probabilistic program concisely represents the concepts required to specify inferences from diverse patterns of evidence. In two experiments, we demonstrated a very close fit between our model's predictions and participants' judgements. Our model accurately predicted how people reason with confounded and indirect evidence and how different sources of information are integrated.

7.2 Future Directions

7.2.1 Casting out the devil in the details

The framework developed in this thesis predicts that responsibility attributions are influenced by how critical a person's action was perceived for the outcome and how close the person's action was to being pivotal *ex post*. These general predictions have been supported by a series of experiments (Gerstenberg & Lagnado, 2010; Lagnado et al., accepted; Zultan et al., 2012) and we are confident that they will hold up in other situations as well. As the summary of the main findings of this thesis has shown, the framework serves a valuable organisational function by predicting how different factors influence attributions of responsibility through affecting criticality and pivotality.

However, as so often the case, the devil is in the details. Future research will have to explore exactly how the concepts of criticality and pivotality are linked to the underlying causal structure. As mentioned earlier, we have focused on rather minimal group settings in which the actions of each agent were independent from each other. Exploring situations in which their are interdependencies between agents, such as causal

hierarchies, will lead to refined concepts of criticality and pivotality. It will further be important to investigate how these two components combine to affect attributions. Can we predict from the context of the situation whether responsibility attributions will be more strongly influenced by criticality or pivotality considerations?

Most of the experiments in this thesis have looked at how differences in criticality and pivotality affect responsibility attributions to individuals in groups. What role criticality and pivotality play in situations in which the outcome was brought about by a single individual remains an open question. For example, we might wonder whether a person's responsibility for an outcome reduces in the presence of a second physical cause that overdetermined the outcome.

More generally, it will be crucial to explore in more detail how attributions of responsibility are dependent upon the nature of the causal events of interest. Our framework defines the notions of criticality and pivotality in terms of counterfactual dependence as implied by the causal structure of the situation. Thus, it operates on a high level of abstraction. However, it is clear that responsibility attributions are not only concerned with whether a certain event was present or absent – it matters how the event came about. So far, our framework cannot 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. Delineating the additional factors that are required to capture potential differences in attributions between structurally equivalent situations remains a challenge for future research.

7.2.2 Exploring the relationship between responsibility and regret

Intuitively, responsibility and regret are closely related. Thinking that one's action was responsible for having brought about an outcome seems like an important factor for whether regret will be experienced. Indeed, previous psychological research has argued for a close coupling between responsibility and regret (Connolly & Zeelenberg, 2002; Petrocelli et al., 2011).

In economic theory, regret is defined in terms of the difference between the actual payoff and the (better) counterfactual payoff one could have received if one had made a different choice (Bell, 1982; Loomes & Sugden, 1982). Disappointment, in contrast, is defined as the difference between actual payoff and expected payoff (Bell, 1985). Thus, regret and disappointment are distinguished via the agent's role in bringing about the outcome. Whereas regret requires having made a choice, disappointment doesn't. Research has explored whether people's judgments about what emotions a protagonist is likely to experience when facing a negative outcome follow these predictions. The results have been mixed. While there is some evidence that the relationship between responsibility (or decisional agency) and regret is only weak (Connolly, Ordóñez, & Coughlan, 1997; Ordóñez & Connolly, 2000) others have provided support for a

strong coupling (Zeelenberg, Van Dijk, & Manstead, 1998, 2000).

An important question that has not been explored thus far concerns the relationship between regret and the degree to which a person's action was pivotal for the outcome (see Nicolle, Bach, Frith, & Dolan, 2011, for an empirical demonstration that criticality affects regret). In terms of our general framework, a person's responsibility for an outcome is not simply a function of whether or not a choice was made. The relationship between a person's choice and the outcome is crucial. According to the economic definition, regret is defined in terms of simple counterfactual dependence. A person should only experience regret if the outcome would have been different had she acted differently. This definition implies that a person should experience *no* regret for having made a certain choice when the outcome was overdetermined.

Consider the following simple game. Two agents, A and B, have to choose between two boxes. One of the boxes contains a prize and the other box is empty. The rules are such that if at least one of the agents chooses the black box it will be opened. The white box will only be opened if both agents choose it. Figure 7.1 shows three possible situations in which the agents experience a negative outcome. In Situation 1, both agents chose the black box. In Situation 2, A chose the black and B the white box. Finally, in Situation 3, both agents chose the white box. Let us focus on how much regret A is expected to experience in the different situations.

According to the economic definition of regret, A is predicted to experience *no* regret in Situation 1. As both agents chose the black box, the negative outcome is overdeter-

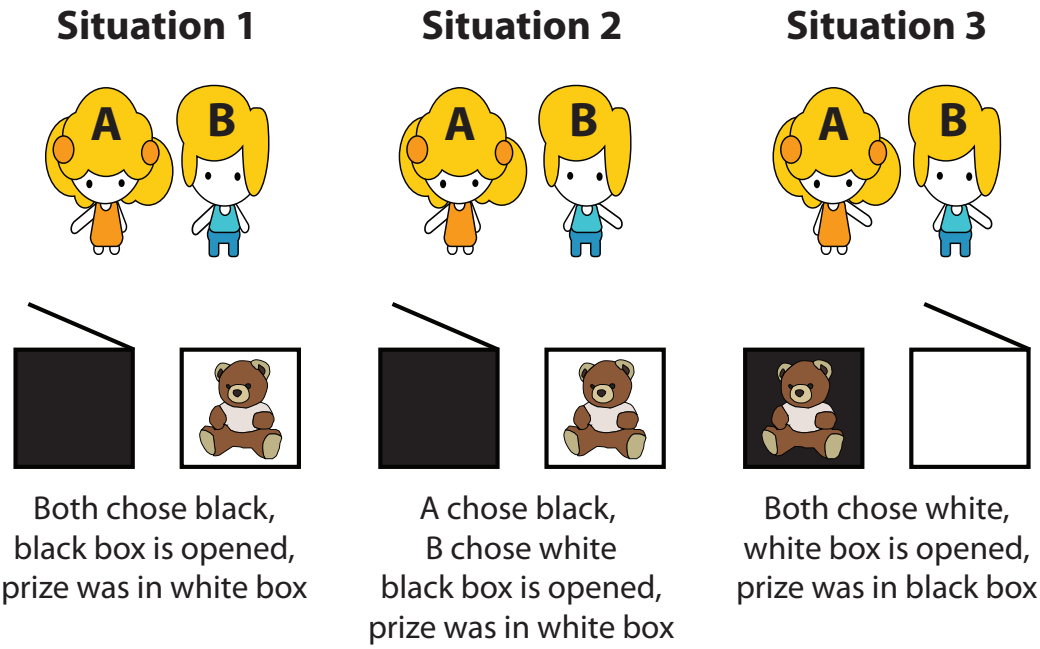


Figure 7.1: Three different situations in the regret game. The black box is opened if at least one child chooses it. The white box is opened only if both children chose it.

mined. Even if A had chosen the white box, the agents still would not have received the prize. In contrast, in Situations 2 and 3, A is predicted to experience regret. In both of these situations, the agents would have received the prize if A had made a different choice.

Our framework, in contrast, would predict that A will also experience regret in Situation 1 (although less so than in Situations 2 and 3). A's action would have made a difference in the possible world in which B had chosen differently. Intuitively, it will also make a difference whether A and B chose the same versus different options. Accordingly, A will feel more regret in Situation 2 compared to Situation 3. Indeed, recent research suggests that people's emotional responses in situations that are conducive to regret are not only dependent on their own choices and payoffs but also on the choices and payoffs of other social agents (see Bault, Coricelli, & Rustichini, 2008). Looking at this issue from a responsibility attribution perspective, B would be predicted to blame A more for the negative outcome in Situation 2 compared to Situation 3.

In sum, I believe that our framework for understanding responsibility attributions in terms of counterfactuals over different possible worlds can help to develop a richer theory of how people's emotional experiences are grounded in their causal understanding of the situation (cf. Weiner, 1995).

7.2.3 Bringing it all together

In this thesis, I have provided a wealth of evidence that bears on the question of how people attribute responsibility. Given the flexibility of our framework, I believe that it is in a good position to integrate different factors that have been shown to affect responsibility attributions in a principled way.

As outlined above, one major advantage of defining causal attributions in terms of counterfactuals over probabilistic generative models is the generality of this approach. Consider a road accident involving Mr Red and Ms Blue (see Figure 7.2). Red was speeding and under the influence of alcohol while Blue had forgotten to turn on the headlights. When answering the question to what extent each person is responsible for the accident we need to be able to consider both counterfactuals on physical variables (e.g. (1) a car's speed) as well as psychological variables (e.g. (2) being drunk, (3) forgetting something). If, for example, we believe that the accident would have happened even if Blue's headlights had been on, we might downweigh the importance of this factor for our responsibility attribution. To what degree we consider Red's drunkenness as an important factor will also depend on our estimate of how the consumption of alcohol increases the risk of an accident (e.g. through an increased reaction time). As this example illustrates, a comprehensive model of responsibility attribution has to draw on both people's intuitive understanding of physics as well as their psychological understanding of how people work.

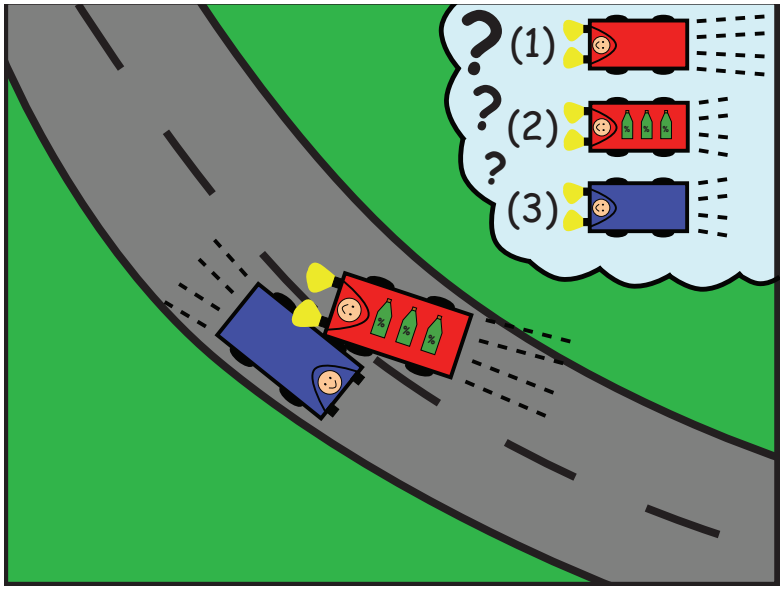


Figure 7.2: A road accident with relevant counterfactuals in the top right.

In Chapter 2, we have seen that most research thus far has been unable to find strong dissociations between judgments of causality and responsibility. One viable possibility is that this lack of a finding is due in part to the research methods that were being used. By using more dynamic stimuli (rather than mere descriptions) in which intentionally acting agents are directly observed in their physical environment, we might be able to generate situations in which judgments of causation and blame come apart.

As the road accident example mentioned above illustrates, most settings in which people make attributions contain information about both physical and psychological factors. It would be interesting to explore whether differences between causal and responsibility attributions in such situations can be modelled in terms of a differential consideration of counterfactuals. For example, the degree to which a driver will be judged to have *caused* the accident might be mostly sensitive to counterfactual considerations over physical variables (e.g. whether the accident would have been prevented if the car had been going slower). In contrast, the degree to which the driver will be *blamed* might be related to counterfactuals defined over the mental states of the driver (e.g. what would have happened if the driver had reacted more quickly).

7.3 Implications

In the introduction to this thesis, I reported the story of politician McDonald whose wife did not vote. Since the election ended in a tie, McDonald would have won if only his wife had managed to vote. McDonald did not blame his wife. He had told her that a single vote won't probably make any difference. Although I mentioned that this thesis would not be concerned primarily with what motivates people to vote, the developed

responsibility framework suggests that there is a close relationship between responsibility and motivation (cf. Weiner, 1995).

A person's motivation to contribute towards a group effort should increase to the extent to which she feels prospectively responsible for the outcome (cf. De Cremer & van Dijk, 2002; Kerr, 1983). More generally, the notions of criticality and pivotality as developed in this thesis can be used to analyse the structure of complex systems or institutions. On the one hand, environments should encourage individuals' sense of criticality for the collective outcome. 'Conjunctive' environments generally increase the motivation to put in effort. However, on the other hand, environments should also exhibit robustness. That is, we would not want the outcome of the whole collective to be dependent upon the actions of a single individual. Thus, environments also require disjunctive structures that reduce the chances of a single individual's action being pivotal. Hence, in order to work well, institutions should establish environments in which criticality and pivotality are balanced. For example, through a distribution of tasks to different groups, an environment can be generated in which individuals feel critical for the outcome in their group task but the institution as a whole might not be dependent on the successful performance of each group.

The recent financial crisis illustrates what can happen in imbalanced institutional structures. Indeed, some have argued that the financial crisis was first and foremost a crisis of responsibility (Ferguson, 2010). Bankers were accused for their greed as well as their lack of responsibility. The framework developed in this thesis highlights what structural aspects are likely to be conducive for such behaviour. Rather than merely trying to explain the crisis in terms of the individual banker's greedy disposition it might be more fruitful to focus on how the institutional environment could be changed. Creating structures which foster a sense of collective responsibility may help to decrease the chances of future crises.

Another implication of the framework developed in this thesis is its application to legal reasoning (cf. Fenton et al., 2012; Lagnado, 2011b; Lagnado et al., 2012). It rarely happens that a crime can be attributed to a single cause. More commonly, there are a multitude of factors that jointly culminate in the criminal act. In such situations, the question of how to apportion punishment between the multiple parties involved is pressing. On several occasions, I have highlighted the close ties between our framework and ideas developed by legal theorists (Cane, 2002; Hart, 2008; Hart & Honoré, 1959/1985; Moore, 2009). In particular, our account resonates well with the importance of the notion of prospective responsibility in legal thinking. Indeed, Cane (2002, p. 35) argues that 'historic [i.e. retrospective] responsibility finds its role and meaning only in responding to nonfulfillment of prospective responsibilities'. I believe that our framework has the potential to produce insights that are of value to legal theorists. Our work can contribute to a theoretical refinement of the notions of prospective and retrospective responsibility and investigate exactly how they are interrelated. Furthermore, we can

explore empirically when people's intuitive attributions of responsibility are in line with legal practice and when they diverge.

7.4 Conclusion

As Aldous Huxley's quote at the beginning of this chapter illustrates, we are often inclined to come up with theories that explain what happens in the world – especially for things that are of relevance to us. He might also be right in saying that we are often incapable of finding correct explanations. In this thesis, I have developed a framework that aims to explain how people attribute responsibility. I am almost certain that the developed framework is not a 'correct explanation' (it is certainly not a complete one). However, I believe that it is a good first step towards one. Highlighting the close interrelations between responsibility, causality and counterfactuals resonates well with the core themes that were brought to scientific attention by the founding father (and sons) of attribution theory (Heider, 1958; Kelley, 1967; Weiner, Heckhausen, Meyer, & Cook, 1972). Furthermore, our framework combines key insights from a range of scientific disciplines such as psychology, philosophy, computer science and law (e.g. Hart & Honoré, 1959/1985; Pearl, 2000; Spellman, 1997; Woodward, 2003). Not only will this allow us to share the blame if things go wrong, but such a collaborative effort will hopefully decrease the chances of us acting like lunatics and instead make a difference for the better.

References

Ajzen, I. (1971). Attribution of dispositions to an actor: Effects of perceived decision freedom and behavioral utilities. *Journal of Personality and Social Psychology*, *18*(2), 144–156.

Ajzen, I., & Fishbein, M. (1975). A bayesian analysis of attribution processes. *Psychological Bulletin*, *82*(2), 261–277.

Ajzen, I., & Fishbein, M. (1983). Relevance and availability in the attribution process. In J. M. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Advances in experimental social psychology* (pp. 63–89). New York: Academic Press.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378.

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

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin*, *34*(10), 1371–1381.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696.

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365.

Au, W. T. (2004). Criticality and environmental uncertainty in step-level public goods dilemmas. *Group Dynamics: Theory, Research, and Practice*, *8*(1), 40–61.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Austin, TX: : Cognitive Science Society.

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.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Bault, N., Coricelli, G., & Rustichini, A. (2008). Interdependent utilities: how social ranking affects choice behavior. *PLoS One*, *3*(10), e3477.

Beebee, H. (2004). Causing and nothingness. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 291–308). MA: MIT Press Cambridge.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981.

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations research*, *33*(1), 1–27.

Berkowitz, L. (1972). Social norms, feelings, and other factors affecting helping and altruism. *Advances in Experimental Social Psychology*, *6*, 63–108.

Berkowitz, L., & Daniels, L. R. (1963). Responsibility and dependency. *The Journal of Abnormal and Social Psychology*, *66*(5), 429-436.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193.

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology*, *36*(1), 56–71.

Brandts, J., & Charness, G. (2003). Truth or consequences: An experiment. *Management Science*, *49*(1), 116–130.

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

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, *6*(10), 426–431.

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

CapelloIndex.com. (2012). http://www.capelloindex.com/.

Cartwright, N. (1995). False idealisation: A philosophical threat to scientific method. *Philosophical Studies*, *77*(2), 339–352.

Charness, G., & Levine, D. I. (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal*, *117*(522), 1051–1072.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

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.

Coffee, J. C. (1981). "No soul to damn: No body to kick": An unscandalized inquiry into the problem of corporate punishment. *Michigan Law Review*, *79*(3), 386–459.

Cohen, L. J. (1981). Who is starving whom? *Theoria*, *47*(2), 65–81.

Connolly, T., Ordóñez, L. D., & Coughlan, R. (1997). Regret and responsibility in the evaluation of decision outcomes. *Organizational behavior and human decision processes*, *70*(1), 73–85.

Connolly, T., & Zeelenberg, M. (2002). Regret in decision making. *Current Directions in Psychological Science*, *11*(6), 212–216.

Cooper, D. E. (1968). Collective responsibility. *Philosophy*, *43*(165), 258–268.

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision*, *12*(11), 1–19.

Critchlow, B. (1985). The blame in the bottle. *Personality and Social Psychology Bulletin*, *11*(3), 258–274.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, *4*(8), e6699.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075.

Danks, D., Rose, D., & Machery, E. (in press). Demoralizing causation. *Philosophy and Phenomenological Research*.

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.

Darley, J. M., & Zanna, M. P. (1982). Making moral judgments: Certain culturally transmitted excuses are generally believed to absolve people of blame for harming others. *American Scientist*, *70*(5), 515–521.

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

De Cremer, D., & Van Lange, P. A. M. (2001). Why prosocials exhibit greater cooperation than proselfs: The roles of social responsibility and reciprocity. *European Journal of Personality*, *15*(1), 5–18.

DeLucia, P., & Liddell, G. (1998). Cognitive motion extrapolation and cognitive clocking in prediction motion tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 901–914.

Dennett, D. C. (1989). *The intentional stance*. The MIT press.

Dickson, V., & Theobald, J. (2011). Criminal law. In *Graduate diploma in law.* Guildford: The College of Law.

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

Downs, A. (1957). *An economic theory of democracy.* New York: Harper & Row.

Driver, J. (2008). Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: intuition and diversity* (Vol. 2). The MIT Press.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*, 268–298.

Edgington, D. (2011). Causation first: Why causation is prior to counterfactuals. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology.* Oxford: Oxford University Press.

Edlin, A. S., Gelman, A., & Kaplan, N. (2007). Voting as a rational choice: why and how people vote to improve the well-being of others. *Rationality and Society*, *19*(3), 293–314.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, *41*(1), 20–26.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness: Intentions matter. *Games and Economic Behavior*, *62*, 287–303.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*, 293–315.

Feather, N. T., & Simon, J. G. (1971). Attribution of responsibility and valence of outcome in relation to initial confidence and success and failure of self and other. *Journal of Personality and Social Psychology*, *18*(2), 173–188.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*, 817–868.

Feigenson, N., Park, J., & Salovey, P. (1997). Effect of blameworthiness and outcome severity on attributions of responsibility and damage awards in comparative negligence cases. *Law and Human Behavior*, *21*(6), 597–617.

Feinberg, J. (1968). Collective responsibility. *The Journal of Philosophy*, *65*(21), 674–688.

Feldman, R. S., & Rosen, F. P. (1978). Diffusion of responsibility in crime, punishment, and other adversity. *Law and Human Behavior*, *2*(4), 313–322.

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

Fenton, N., Neil, M., & Lagnado, D. A. (2012). A general structure for legal arguments about evidence using bayesian networks. *Cognitive Science*, 1-42.

Ferguson, C. D. (2010). *Inside Job [Motion Picture]*. USA: Sony Pictures Classics.

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.

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

Finke, R. A., & Pinker, S. (1982). Spontaneous imagery scanning in mental extrapolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(2), 142–147.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., ... Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517–537.

Fischhoff, B. (1975). Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal ol Experimental Psychology: Human Perception and Performance*, *1*, 288–299.

Fishbein, M., & Ajzen, I. (1973). Attribution of responsibility: A theoretical note. *Journal of Experimental Social Psychology*, *9*(2), 148–153.

Fleishman, J. A. (1988). The effects of decision framing and others' behavior on cooperation in a social dilemma. *Journal of Conflict Resolution*, *32*(1), 162–180.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Forsyth, D. R., & Kelley, K. N. (1994). Attribution in groups estimations of personal contributions to collective endeavors. *Small Group Research*, *25*(3), 367–383.

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

French, P. A. (1984). *Collective and corporate responsibility*. New York: Columbia University Press.

Frieze, I., & Weiner, B. (1971). Cue utilization and attributional judgments for success and failure. *Journal of Personality*, *39*(4), 591–605.

Gailey, J. A., & Falk, R. F. (2008). Attribution of responsibility as a multidimensional concept. *Sociological Spectrum*, *28*(6), 659–680.

Gerstenberg, T. (2009). The allocation of responsibility amongst multiple causes. *Un-*

published MSc thesis.

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 720–725). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Goodman, N. D. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). 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.

Gerstenberg, T., Lagnado, D. A., & Kareev, Y. (2010). The dice are cast: The role of intended versus actual contributions in responsibility attribution. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1697–1702). Austin, TX: Cognitive Science Society.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38.

Gilbert, M. (1997). Group wrongs and guilt feelings. *The Journal of Ethics*, *1*(1), 65–84.

Gilbert, M. (2006). Who's to blame? collective moral responsibility and its implications for group members. *Midwest studies in philosophy*, *30*(1), 94–114.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1-3), 111–133.

Glymour, C. (2007). Statistical jokes and social effects: Intervention and invariance in causal relations. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 294–300). Oxford: Oxford University Press.

Godfrey-Smith, P. (2010). Causal pluralism. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *Oxford handbook of causation* (pp. 326–337). Oxford University Press.

Goodman, N. (1983). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social

reasoning in causal learning. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Uncertainty in Artificial Intelligence.*

Goodman, N. D., & Stuhlmüller, A. (2012). Knowledge and implicature: Modeling language understanding as social cognition. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Goodman, N. D., & Tenenbaum, J. B. (in prep). *The probabilistic language of throught.*

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2010). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110.

Gopnik, A., & Wellman, H. (in press). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*.

Green, E. (1967). The reasonable man: Legal fiction or psychosocial reality? *Law & Society Review*, *2*, 241–258.

Griffiths, T. L. (2005). Causes, coincidences, and theories. *Unpublished doctoral dissertation*.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition*, *117*(2), 139–150.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry*, *52*(5), 449–466.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388.

Gweon, H., & Schulz, L. E. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.),

Causation and counterfactuals. MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, *132*, 109–136.

Halpern, J. Y. (2008). Defaults and normality in causal structures. In *Proceedings of the 11th Conference on Principles of Knowledge Representation and Reasoning* (pp. 198–208).

Halpern, J. Y., & Hitchcock, C. (2011). Actual causation and the art of modeling. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (pp. 316–328). College Publications.

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.

Hamilton, V. L. (1980). Intuitive psychologist or intuitive lawyer? Alternative models of the attribution process. *Journal of Personality and Social Psychology*, *39*(5), 767–772.

Hamilton, V. L. (1986). Chains of command: Responsibility attribution in hierarchies. *Journal of Applied Social Psychology*, *16*(2), 118–138.

Hamrick, J., Battaglia, P., & Tenenbaum, J. (2011). Internal physics models guide probabilistic judgments about object dynamics. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hardin, G. (1968). The tragedy of the commons. *Science*, *126*, 1243–1248.

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.

Harvey, M. D., & Rule, B. G. (1978). Moral evaluations and judgments of responsibility. *Personality and Social Psychology Bulletin*, *4*(4), 583–588.

Hegarty, M. (1992). Mental animation: inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1084–1102.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Hegarty, M., Just, M. A., & Morrison, I. R. (1988). Mental models of mechanical systems: Individual differences in qualitative and quantitative reasoning. *Cognitive Psychology*, *20*(2), 191–236.

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

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The*

American Journal of Psychology, *57*(2), 243–259.

Hertel, G., Kerr, N. L., & Messé, L. A. (2000). Motivation gains in performance groups: Paradigmatic and theoretical developments on the Köhler effect. *Journal of Personality and Social Psychology*, *79*(4), 580–601.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.

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.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.

Hitchcock, C. (1995). Salmon on explanatory relevance. *Philosophy of Science*, *62*(2), 304–320.

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

Hitchcock, C. (2001b). A tale of two effects. *The Philosophical Review*, *110*(3), 361–396.

Hitchcock, C. (2009). Structural equations and causation: six counterexamples. *Philosophical Studies*, *144*(3), 391–401.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

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

Hobbes, T. (1982/1651). *Leviathan*. Scolar Press.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press, USA.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Hubbard, T. L. (1995). Cognitive representation of motion: evidence for friction and gravity analogues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 241–254.

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*(5), 822–851.

Huck, S. (1999). Responder behavior in ultimatum offer games with incomplete information. *Journal of Economic Psychology*, *20*(2), 183–206.

Huebner, B. (2011). Critiquing empirical moral psychology. *Philosophy of the Social Sciences*, *41*(1), 50–83.

Hume, D. (1748/1975). *An enquiry concerning human understanding*. Oxford University Press.

Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition*, *40*(8), 1387–1401.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses

and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. *Advances in Experimental Social Psychology*, *2*, 219–266.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior*. Morristown, NJ: General Learning Press.

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

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge, England: Cambridge University Press.

Kahneman, D., & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, *59*(6), 1101–1110.

Kanazawa, S. (1992). Outcome or expectancy? antecedent of spontaneous causal attribution. *Personality and Social Psychology Bulletin*, *18*(6), 659–668.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Kelley, H. H. (1972). *Causal schemata and the attribution process*. New York: General Learning Press.

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

Kelley, H. H. (1983). Perceived causal structures. In J. M. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 343–369). New York: Academic Press.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*(1), 457–501.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology*, *45*(4), 819–828.

Kerr, N. L. (1989). Illusions of efficacy: The effects of group size on perceived efficacy in social dilemmas. *Journal of Experimental Social Psychology*, *25*(4), 287–313.

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. (1995). Norms in social dilemmas. In *Social dilemmas: Perspectives on individuals and groups* (pp. 31–47). Praeger Westport, CT.

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. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–324.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, *130*(2), 203–231.

Knobe, J. (2009). Folk judgments of causation. *Studies In History and Philosophy of Science Part A*, *40*(2), 238–242.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–365.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: intuition and diversity* (Vol. 2). The MIT Press.

Knobe, J., & Nichols, S. (2008). *Experimental philosophy.* USA: Oxford University Press.

Kohlberg, L. (1983). *The philosophy of moral development.* New York: Harper & Row.

Kruger, J., & Gilovich, T. (1999). "Naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology*, *76*(5), 743–753.

Kun, A., & Weiner, B. (1973). Necessary versus sufficient causal schemata for success and failure. *Journal of Research in Personality*, *7*(3), 197–207.

Lagnado, D. A. (2011a). Causal thinking. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 129–149). Oxford: Oxford University Press.

Lagnado, D. A. (2011b). Thinking about evidence. *Proceedings of the British Academy*, *171*, 183–223.

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., Fenton, N., & Neil, M. (2012). Legal idioms: A framework for evidential reasoning. *Argument and Computation*.

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

Lagnado, D. A., Waldmann, M. ., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy,*

and computation (pp. 154–172). Oxford University Press.

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

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, *89*(2), 308–324.

Lazarus, R. S. (1966). *Psychological stress and the coping process.* New York: McGraw-Hill.

Leddo, J., Abelson, R. P., & Gross, P. H. (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology*, *47*(5), 933–943.

LeDoux, J. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, *23*(1), 155–184.

Lewin, K. (1936). *Principles of topological psychology.* New York: McGraw-Hill.

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

Lewis, H. D. (1948). Collective responsibility. *Philosophy*, *23*(84), 3–18.

Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, *109*(3), 456–471.

Livengood, J. (2011). Actual causation and simple voting scenarios. *Noûs*, 1–33.

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

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824.

Lucas, C. G., & Kemp, C. (2012). A unified theory of counterfactuals. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Luce, R. (1959). *Individual choice behavior: A theoretical analysis.* John Wiley.

Mackie, J. L. (1974). *The cement of the universe.* Oxford: Clarendon Press.

Mäkelä, P. (2007). Collective agents and moral responsibility. *Journal of Social Philosophy*, *38*(3), 456–468.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*(6), 895–919.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

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.

Manning, R., Levine, M., & Collins, A. (2007). The kitty genovese murder and the social psychology of helping: The parable of the 38 witnesses. *American Psychologist*,

62(6), 555–562.

Mao, W., & Gratch, J. (2005). Social causality and responsibility: Modeling and evaluation. *Intelligent Virtual Agents*, 191–204.

Markman, K. D., & Tetlock, P. E. (2000). 'i couldn't have known': Accountability, foreseeability and counterfactual denials of responsibility. *British Journal of Social Psychology*, *39*(3), 313–325.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* W. H. Freeman.

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

Mathiesen, K. (2006). We're all in this together: Responsibility of collective agents and their members. *Midwest Studies in Philosophy*, *30*(1), 240–255.

May, L. (1993). *Sharing responsibility.* Chicago: University of Chicago Press.

McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, *22*(2), 171–193.

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, *52*, 267–275.

McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine, part i. *Communications of the ACM*, *3*(4), 184–195.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naíve beliefs about the motion of objects. *Science*, *210*(4474), 1138–1141.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649.

McClure, J. (1998). Discounting causes of behavior: Are two reasons better than one? *Journal of Personality and Social Psychology*, *74*(1), 7–20.

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.

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, *123*(1), 125–148.

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *Open*

Psychology Journal, *3*, 119–135.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249–264.

Mele, A. R. (1997). *The philosophy of action*. Oxford University Press.

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, *130*(5), 711–47.

Michotte, A. (1946/1963). *The perception of causality*. Basic Books.

Mill, J. S. (1898). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Longmans, Green.

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.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225.

Mitchell, T. R., & Kalb, L. S. (1981). Effects of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology*, *66*, 604–612.

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

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355.

Mynatt, C., & Sherman, S. J. (1975). Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, *32*(6), 1111–1118.

Nadelhoffer, T. (2005). Skill, luck, control, and intentional action. *Philosophical Psychology*, *18*(3), 341–352.

Nagel, T. (1979). Moral luck. In *Mortal questions* (pp. 24–38). Cambridge University Press Cambridge.

Narveson, J. (2002). Collective responsibility. *The Journal of Ethics*, *6*(2), 179–198.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nicolle, A., Bach, D. R., Frith, C., & Dolan, R. J. (2011). Amygdala involvement in self-blame regret. *Social Neuroscience*, *6*(2), 178–189.

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, inten-

tion, and outcome on children's moral judgments. *Journal of Experimental Child Psychology*, *104*(4), 382–397.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* USA: Oxford University Press.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, *46*(8), 1423–1437.

Ordóñez, L. D., & Connolly, T. (2000). Regret and responsibility: A reply to Zeelenberg et al. (1998). *Organizational Behavior and Human Decision Processes*, *81*(1), 132–142.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann.

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

Pearl, J. (2008). Review of N. Cartwright "Hunting Causes and Using Them". *Economics and Philosophy*, *26*, 69–77.

Pearl, J. (2011). The structural theory of causation. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences.* Oxford University Press.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46.

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.

Pettit, P. (2007). Responsibility incorporated. *Ethics*, *117*(2), 171–201.

Piaget, J. (1932). *The moral judgement of the child.* London: Kegan Paul, Trench, Trubner and Co.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–99.

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, *36*(1), 11–46.

Rantilla, A. K. (2000). Collective task responsibility allocation revisiting the group-serving bias. *Small group research*, *31*(6), 739–766.

Rapoport, A. (1985). Provision of public goods and the mcs experimental paradigm. *The American Political Science Review*, *79*(1), 148–155.

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

Rapoport, A., Bornstein, G., & Erev, I. (1989). Intergroup competition for public goods: Effects of unequal resources and relative group size. *Journal of Personality and*

Social Psychology, *56*(5), 748–756.

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

Richardson, H. L., Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2012). The development of joint belief-desire inferences. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.

Rips, L. J. (2011). Causation from perception. *Perspectives on Psychological Science*, *6*(1), 77–97.

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.

Roessler, J., Lerman, H., & Eilan, N. (2011). *Perception, causation, and objectivity.* Oxford: Oxford University Press.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*(3), 322–336.

Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing.* MIT Press.

Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*(2), 297–312.

Samuelson, P. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Sanders, J., Hamilton, V., Denisovsky, G., Kato, N., Kawai, M., Kozyreva, P., ... Tokoro, K. (2006). Distributing responsibility for wrongdoing inside corporate hierarchies: Public judgments in three societies. *Law & Social Inquiry*, *21*(4), 815–855.

Sanna, L. J., & Turley, K. J. (1996). Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin*, *22*(9), 906–919.

Sartorio, C. (2005). Causes as difference-makers. *Philosophical Studies*, *123*(1), 71–96.

Sartorio, C. (2007). Causation and responsibility. *Philosophy Compass*, *2*(5), 749–765.

Savitsky, K., Van Boven, L., Epley, N., & Wight, W. M. (2005). The unpacking effect in allocations of responsibility for group tasks. *Journal of Experimental Social Psychology*, *41*(5), 447–457.

Schächtele, S., Gerstenberg, T., & Lagnado, D. A. (2011). Beyond outcomes: The influence of intentions and deception. In L. Carlson, C. Hölscher, & T. Shipley

(Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1860–1865). Austin, TX: Cognitive Science Society.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399.

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.

Schlenker, B. R., & Miller, R. S. (1977). Egocentrism in groups: Self-serving biases or logical information processing? *Journal of Personality and Social Psychology*, *35*(10), 755–764.

Schlottmann, A. (2000). Is perception of causality modular? *Trends in Cognitive Sciences*, *4*(12), 441–441.

Schlottmann, A., & Anderson, N. H. (1993). An information integration approach to phenomenal causality. *Memory & Cognition*, *21*(6), 785–801.

Schlottmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *The Quarterly Journal of Experimental Psychology*, *44*(2), 321–342.

Scholl, B. J. (2008). Two kinds of experimental philosophy (and their methodological dangers). In *SPP Workshop on Experimental Philosophy.* Philadelphia, PA: University of Pennsylvania.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung.* Tübingen: JCB Mohr.

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 181–190.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science*, *23*(10), 1264–1270.

Shanks, D. R. (1995). *The psychology of associative learning.* Cambridge, England: Cambridge University Press.

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.

Shaw, M. E., & Sulzer, J. L. (1964). An empirical test of Heider's levels in attribution of responsibility. *The Journal of Abnormal and Social Psychology*, *69*(1), 39–46.

Sheehy, P. (2003). Social groups, explanation and ontological holism. *Philosophical Papers*, *32*(2), 193–224.

Shultz, T. R. (1986). *A computational model of blaming.* (Presented at the Annual Meetting of the Society of Experimental Social Psycholgy, Phoenix)

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

Shultz, T. R., & Schleifer, M. (1983). Towards a refinement of attribution concepts. In J. M. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 37–62). London: Academic Press.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science*, *13*(3), 238–253.

Sinnott-Armstrong, W. (2008). *Moral psychology: The cognitive science of morality: intuition and diversity* (Vol. 2). The MIT Press.

Skyrms, B. (1984). Epr: Lessons for metaphysics. *Midwest Studies in Philosophy*, *9*(1), 245–255.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press, USA.

Sloman, S. A., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, *33*(1), 21–50.

Sloman, S. A., & Lagnado, D. A. (2005). Do we 'do'? *Cognitive Science*, *29*(1), 5–39.

Smiley, M. (2011). Collective responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2011 ed.).

Smith, K. A., & Vul, E. (2012). Sources of uncertainty in intuitive physics. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Sousa, P. (2009). A cognitive approach to moral responsibility: The case of a failed attempt to kill. *Journal of Cognition and Culture*, *9*(3), 171–194.

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.

Stanca, L. (2010). How to be kind? outcomes versus intentions as determinants of fairness. *Economics letters*, *106*(1), 19–21.

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

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

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

Stuhlmüller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning structured generative concepts. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Suppes, P. (1970). *A probabilistic theory of causation.* North-Holland.

Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*(1), 69–78.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction.* Cambridge University Press.

Sverdlik, S. (1987). Collective responsibility. *Philosophical Studies*, *51*(1), 61–76.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Teigen, K. H., & Brun, W. (2011). Responsibility is divisible by two, but not by three or four: Judgments of responsibility in dyads and groups. *Social Cognition*, *29*(1), 15–42.

Teigen, K. H., Kanten, A. B., & Terum, J. A. (2011). Going to the other extreme: Counterfactual thinking leads to polarised judgements. *Thinking & Reasoning*, *17*(1), 1–29.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In *Causal learning: Psychology, philosophy, and computation.* Oxford University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471.

Thompson, J. (2006). Collective responsibility for historic injustices. *Midwest Studies in Philosophy*, *30*(1), 154–167.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Tollefsen, D. (2006). The rationality of collective guilt. *Midwest Studies in Philosophy*, *30*(1), 222–239.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution.

Psychological Review, *93*(3), 239–257.

Tyler, T. R., & Devinitz, V. (1981). Self-serving bias in the attribution of responsibility: Cognitive versus motivational explanations. *Journal of Experimental Social Psychology*, *17*(4), 408–416.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *17*, 455-480.

Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (Vol. 22).

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236.

Walker, C. M., & Gopnik, A. (forthcoming). Causality and imagination.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, *26*(1), 21–52.

Walsh, V., & Kulikowski, J. J. (1998). *Perceptual constancy: Why things look as they do*. Cambridge University Pres.

Waytz, A., & Young, L. (2012). The group-member mind trade-off. *Psychological Science*, *23*(1), 77–85.

Weber, M. (1914/1978). *Economy and society* (Vol. 1). Berkeley: University of California Press.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548.

Weiner, B. (1991). Metaphors in motivation and attribution. *American Psychologist*, *46*(9), 921–930.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: The Guilford Press.

Weiner, B., Frieze, I., Kukla, A., Reed, L., Rest, S., & Rosenbaum, R. (1971). *Perceiving the causes of success and failure.* New York: General Learning Press.

Weiner, B., Heckhausen, H., Meyer, W. U., & Cook, R. E. (1972). Causal ascriptions and achievement behavior: A conceptual analysis of effort and reanalysis of locus of control. *Journal of Personality and Social Psychology*, *21*(2), 239–148.

Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation.

Journal of Personality and Social Psychology, *15*(1), 1–20.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.

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

Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology*, *53*(3), 421–430.

White, P. A. (2007). Impressions of force in visual perception of collision events: A test of the causal asymmetry hypothesis. *Psychonomic Bulletin & Review*, *14*(4), 647–652.

Wikipedia. (2012). *Great Heck Rail Crash — Wikipedia, The Free Encyclopedia.* Retrieved from `http://en.wikipedia.org/w/index.php?title=Great_Heck_Rail_Crash&oldid=508217075` (accessed 5-September-2012)

Williams, B. (1981). *Moral luck: Philosophical papers, 1973-1980.* Cambridge University Press.

Williamson, J. (2006). Causal pluralism versus epistemic causality. *Philosophica*, *77*, 69–96.

Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*(1), 1–48.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

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

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, *47*(3), 276–332.

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.

Woodward, J. (2011a). Causal perception and causal cognition. In J. Roessler, H. Lerman, & N. Eilan (Eds.), *Perception, causation, and objectivity.* Oxford: Oxford University Press.

Woodward, J. (2011b). Psychological studies of causal and counterfactual reasoning. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology.* Oxford: Oxford University Press.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301.

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

Zeelenberg, M., Van Dijk, W. W., & Manstead, A. S. R. (1998). Reconsidering the relation between regret and responsibility. *Organizational Behavior and Human Decision Processes*, *74*(3), 254–272.

Zeelenberg, M., Van Dijk, W. W., & Manstead, A. S. R. (2000). Regret and responsibility resolved? Evaluating Ordóñez and Connolly's (2000) conclusions. *Organizational Behavior and Human Decision Processes*, *81*(1), 143–154.

Zhou, J., Huang, X., Jin, X., Liang, J., Shui, R., & Shen, M. (in press). Perceived causalities of physical events are influenced by social cues. *Journal of Experimental Psychology: Human Perception and Performance*.

Zimmerman, M. (1985). Sharing responsibility. *American Philosophical Quarterly*, *22*(2), 115–122.

Zimmerman, M. (2001). Responsibility. In L. Becker (Ed.), *Encyclopedia of ethics* (pp. 1089–1095). New York: Garland.

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