Mind, Matter, Morals - The Epistemic Condition in Causal Judgment

Lara Kirfel

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Experimental Psychology

University College London

October 25, 2021

I, Lara Kirfel, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In this thesis, we explore the role of theory of mind, broadly construed, in people's causal reasoning and inferences from causal explanations. While research in causal cognition has acknowledged the influence of agents' knowledge states on how causal they are judged, the theoretical implications of these findings remain unclear. This thesis aims to provide an empirical and theoretical account of the essential function of epistemic states in causal thinking and inference.

In chapter 2, we demonstrate how epistemic states mediate a prominent finding in causal cognition research — people's preference for deviant causal agents. Various studies in causal cognition research find that people have the tendency to attribute increased causality to atypical actions. In a series of experiments, we find that this abnormal causal preference is driven by the epistemic states of the causal agents. In chapter 3, we show more generally what role causal agents' epistemic states and epistemic actions play for people's causal judgments. We develop and test an account that integrates epistemic states into counterfactual theories of causation: agents' epistemic states influence the target of intervention in people's counterfactual reasoning. In chapter 4, we investigate whether other people's epistemic states not only influence causal judgments, but also play a role for the inferences we draw from their causal explanations.

In sum, this thesis shows how people take into account the mental states of an agent in both causal judgment and inference. We situate these findings in the broader context of causal and counterfactual theories.

Impact Statement

This thesis investigates the intersection of two fundamental abilities of the human mind: the ability to judge about and infer causality (“causal cognition”), and the ability to understand other agents’ minds (“theory of mind”). In a series of studies, I have demonstrated that these two domains are closely intertwined in human cognition. Knowing the causes of things imparts power: We can explain and understand the world around us and change it for the better. When we apply our causal knowledge in the social world, we do so in combination with our theory of mind. Our ability to infer the mental states of others is crucial. We grow up in rich social environments, being socialised through the acquisition of a sophisticated theory of mind of the people around us. In this thesis, I show that our understanding of the causality of agents cannot be separated from our understanding of their minds. People’s theory of mind dictates how they think about alternatives, and the inferences they draw from them.

Contemporary formal frameworks of causality have turned a blind eye on the fundamental role of mental states in agent causal reasoning. At the same time, these causal frameworks are used to inform the ever growing use and development of algorithms to reveal causal patterns in large data sets (Pearl, 2019). In order to identify precise causal factors that can provide the basis for intervention and lead to particular behaviours or outcomes, my work suggests that AI should aim to code for ‘mental’ causes as well. This extension might enable researchers and practitioners to focus on the best mix of interventions for addressing some of today’s most critical issues, from climate change to health care.

With the increasing involvement of algorithmic decision-making in our daily lives, there is already an emerging field of research on normative and psychological theories of responsible machine behaviour (Rahwan et al., 2019). In order to align these theories with our intuitive theories of causality and responsibility, I argue that references to knowledge states of AI systems may be indispensable. Providing a formalisation of the functional roles of some of the core mental states might resolve this problem (Ashton, 2021; Quillien & German, 2021), but also shed further theoretical insight into our mental taxonomy.

Finally, the rise of AI has not only led to the need of causality and causal inference, but also the need for explanation. Explanations that give insight into algorithmic decision-making – "Explainable AI" – are notoriously difficult to develop (Holzinger, 2018). The integration of Theory of Mind into explanations holds the potential of producing high-quality explanations that are tailored to the beliefs of the listener, in the context of the beliefs of the speaker (Shvo, Klassen, & McIlraith, 2020). Endorsing epistemic-based accounts of explanations, as put forward in this thesis, might help AI systems to produce personalised explanations that take into account users world knowledge and belief revision.

As the work in this thesis suggests, incorporating epistemic states is critical for building better AI systems that are capable of making better causal inferences, taking responsible decisions as well as providing good explanations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prologue: Causal Minds

Anyone who has ever done a Google image search for the term "causality", perhaps in order to illustrate the topic in a talk or a paper, will end up with very similar search results. Pictures of colliding billiard balls, falling dominoes or the swinging balls of Newton's cradle demonstrate what is commonly associated with or imagined under the term causality. While the process of causality is indeed difficult to illustrate, these images of causal chains between inanimate objects neglect a special kind of causal factor that we encounter frequently in our daily lives — other human beings.

In fact, the causal powers of humans are enormous. In early 2021, a small group of Reddit users caused a widespread stock market crash ("Reddit Traders in r/wallstreet Shake Up the Stock Market", 2016, February 26). In 2005, Helios flight 522 crashed into a Greek hillside presumably because one man forgot to flip a switch ("In 2005, Helios flight 522 crashed into a Greek hillside. Was it because one man forgot to flip a switch?", 2020, September 19). The bible describes in detail how, because of a single action by Adam and Eve, mankind is caused to live in sin and sufferance (*The Holy Bible*, n.d.). In his "Metaphysics", Aristotle assumes that the primary cause of all motion in the universe must be some kind of being that he conceives of as an "unmoved mover" (Ross et al., 1924). During the Covid-19 pandemic, a few people have caused a whole web of infections around the world. And finally, although the subject of ongoing discussion, the Intergovernmental Panel on

Climate Change just recently confirmed that that global warming is “unequivocally caused by humans” (“Climate crisis ‘unequivocally’ caused by human activities, says IPCC report”, 2021, August 9). Nonetheless, the typical depiction of causation tends to omit the “human-in-the-loop” (X. Wu et al., 2021).

However, not only do we experience other people as causes; the people we interact with are also often the source of our own causal knowledge about the world. The explanations we receive from others might in fact be the most common way in which we acquire our vast knowledge about the causal processes and workings in the world (Fiorella & Mayer, 2014). Be it through parents, peers, teachers or media, it is thanks to other people that we learn how the world works, and most frequently through the medium of explanations. Without others communicating their causal knowledge to us, we would only be able to acquire a fraction of the knowledge about the world that we possess.

Despite the fact that we as human agents frequently act as causes as well as communicate about causality to others, theories of causality and causal cognition have predominantly focused on object causation (Halpern, 2016; Hume, 1748/1975; Lewis, 1973; Pearl, 2009). How causation works, how people select causes and make judgements about causal strength is often studied with inanimate objects. Social causation, i.e. causation by human agents, it is tacitly assumed, works in a similar way. We reason about causes in a certain way, independent of what the entity of the cause is.

In this thesis, we aim to undertake a first step towards bringing back the “human” in causation. More precisely, we aim to study people’s causal reasoning, and in particular, causal reasoning about agents, as well as people’s inferences from causal explanations. In three chapters, we will argue that there is a common theme underlying both how we make as well as draw inferences from causal judgments — the common theme of our theory of mind, the ability to refer, intuit and infer the mental states of others. We will demonstrate across three research projects that agents’ epistemic states play a crucial role, both for how we make causal judgements about them, as well as for the inferences we draw from their causal explanations.

In chapter 1, we show that a prominent pattern in causal cognition — the selection of abnormal causal actions — can actually be explained with reference to the agents' epistemic states. How we and others normally and frequently behave has an impact on the things we can foresee, and in particular causal structures, an impact on foreseeing the consequences of our actions. We demonstrate in four experiments that it is the epistemic factor that is driving people's tendency to attribute more causality to abnormally acting agents.

In chapter 2, we will pose the more general question what role epistemic states play for causal reasoning about agents. In four experiments, we show that an agent's state of ignorance as well as the epistemic conditions of their ignorance influence how causal they are perceived. We find evidence that epistemic states function as points of intervention for people's counterfactual reasoning about the causal scenario, and as such play a fundamental role for how people judge causality.

Next, we want to reverse the research programme conducted so far, and not only ask what factors influence people's causal judgments, but also what factors people infer from someone else's causal explanation. In chapter 3, we probe whether people are able to make systematic inferences about the aforementioned factor – normality –, as well as about causal structure from the causal statements of others. Here again, we will focus on the crucial role of epistemic states, and more precisely, our ability to have a theory of mind, for drawing systematic inferences from other people's judgements and explanations.

Finally, we will summarise the findings of this thesis with a general reflection on the relationship between theory of mind and causality.

1.2 Causation – A Primer

Before being able to approximate the question how we make causal judgments about social agents, it is crucial to take a broader look at how people reason about causality in general. Being able to identify causes and causal relationships in the world lays the cognitive foundation for a variety of abilities – to explain (Lombrozo, 2010), to make predictions (Einhorn & Hogarth, 1985), to intervene in and control the environment (Woodward, 2007), and finally to make a variety of normative and moral judgments (Malle & Nelson, 2003). As psychologists, our understanding of causation is crucial: In order to understand how humans learn, reason and do all the above, we first and foremost need to understand how they come to learn and reason about causality.

Causal theory distinguishes between *General causation*, the causal regularities between classes of causes and effects ("Smoking causes lung cancer"), and *Singular causation*, the causal relationship between a particular instance of a cause and effect ("Ben's smoking over decades has caused him to develop lung cancer.") (Danks, 2017). Throughout this thesis, the main focus of investigation will be on people's judgments about singular causation. However, singular causation judgments cannot be studied without reference to theories in general causation — judgments about singular causation require knowledge about general causal relations, as well as knowledge about details of the particular causal situation (Danks, 2017; Hitchcock, 1995). But what constitutes a causal relationship? What makes C a cause of E? In causal theory, it is common to distinguish between two major frameworks: *dependency theories* and *process theories*. In the following, we will briefly sketch each of these accounts, discuss their respective advantages and shortcomings as well as further developments.

1.2.1 Dependence theories of causation

According to dependence theories of causation, C is a cause of E if E is in some way dependent on C. On the one hand, this dependence relationship has been spelled out in terms of temporal regularities (Hume, 1748/1975; Mill, 1875). Hume famously argued that C is a cause of E if E was regularly followed by C in the past (Hume, 1748/1975). David Lewis (Lewis, 1973) later criticised Hume and proposed an analysis of causation in terms of counterfactuals. This laid the basis for defining the causal dependence relationship in terms of *difference making*, that is, to what extent a causal candidate makes a difference to the outcome. On the standard counterfactual model of causation (Lewis, 1973, 1979), C is a cause of E if both C and E occur, but E would not have happened in the absence of C. The definition of causation in terms of counterfactual dependence is appealing in its simplicity and intuitiveness, and corresponds to the “but-for” test used in legal liability (Hart & Honoré, 1959/1985), also sometimes referred to as *sine qua non*-condition: But for the act having occurred, the injury or harm would not have happened, rendering the action a proximate cause for the harm (Knobe & Shapiro, 2021).

But what exactly does it mean for a causal factor to make a difference? The notion of difference-making has been formulated and conceptualised in different ways. Interventionist theories specify counterfactuals in terms of hypothetical interventions, often represented in form of a *do*-operator, $do(X = x))$ (Pearl, 2009; Woodward, 2007). Frameworks have cashed out the do-operator differently, e.g. setting a variable to a certain value (Pearl, 2009) or removing the causal candidate from the scene (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2020). However, the basic counterfactual account faces difficulties as soon as no straightforward counterfactual dependence relationship is given. That is the case for example in overdetermination or pre-emption (Collins, 2000; Moore, 1999). The classic scenario to illustrate pre-emption is the following (Hall, Paul, et al., 2003): Billy and Suzy each throw stones at a bottle. Suzy’s stone hits the bottle a few seconds earlier than Billy’s stone and the bottle shatters. Each throw was such that it would have been individually sufficient to shatter the bottle. Was Suzy a cause of the bottle’s shattering? Intuitively, the answer is ‘yes’. Dependence theories, however, struggle to explain causal selection here (Halpern & Hitchcock, 2011). The bottle would still have shattered even if Suzy hadn’t thrown her stone.

Halpern (2016) extend the basic test for counterfactual dependence to different contingencies, i.e. non-actual possible worlds in which certain background variables are different. According to their extension, counterfactual dependence is evaluated both in the actual world as well as relative to certain contingencies. Whether counterfactual dependence holds is hence assessed conditional on certain other interventions on variables in the causal scenario. This allows the account to deal with cases of over-determination, pre-emption and several other classic problems for counterfactual theories. In a counterfactual world in which Billy does not throw the stone, the outcome is counterfactually dependent on Suzy's action (Gerstenberg et al., 2020; Gerstenberg & Lagnado, 2012; Halpern & Hitchcock, 2011).

Dependence theories have also been characterised in probabilistic terms, i.e. C raising the probability that E will happen. If the probability of the effect is lower in the absence of the factor compared to the presence of it – i.e., the factor covaries with the effect (Cheng & Novick, 1991) –, the effect counterfactually depends on the factor which is therefore considered to be causal (Fenton-Glynn, 2017). Dependence theories differ in various ways, but all of them see causes as difference makers and contrast the case in which both the potential causal factor and the effect are present with the counterfactual case in which this causal factor is absent.

In fact, there is plenty of evidence for the influence of counterfactual thinking on causal inferences and judgments (Gerstenberg et al., 2020; Kominsky et al., 2019; Kominsky & Phillips, 2019; Kominsky et al., 2015; Lagnado, 2011; Phillips & Shaw, 2015; Spellman & Gilbert, 2014; Wells & Gavanski, 1989). For example, Gavanski and Wells (1989) used the following scenario in which an outcome counterfactually either did nor did not depend on the presence of a causal factor. In one of their experimental scenarios, a woman died from an allergic reaction after having eaten a meal that had been ordered by her boss. If an alternative meal had not contained the ingredients the woman was allergic to (i.e. counterfactual dependence), the boss is judged as more causal than if the alternative meal had also contained the ingredient and the woman had died anyway (i.e. no counterfactual dependence) (Gavanski & Wells, 1989; Wells & Gavanski, 1989).

1.2.2 Process Theories

According to process theories, C is a cause of E if C and E are connected via a spatio-temporally continuous process that transferred some quantity such as physical force from C to E (Dowe, 2001; Salmon, 1994; Wolff & Shepard, 2013). Several accounts capture the intuition that a cause needs to generate an effect in a different way. Some accounts either emphasise the role of an energetic transfer (Fair, 1979) or some other kind of quantity or force (Dowe, 2001; Wolff & Shepard, 2013)

Process theories have no difficulty in explaining cases such as causal pre-emption and overdetermination, because in each case there is a spatiotemporally continuous process from each cause to the effect (Wolff, 2003). Take again the case of pre-emption. According to process theories, Suzy – but not Billy – is a cause of the shattered bottle because her action is spatiotemporally connected with the shattered bottle, unlike Billy's. However, process theories of causation struggle to accommodate causation by omission (Gerstenberg & Stephan, 2021; McGrath, 2005; Schaffer, 2000; Willemsen, 2018). Under certain circumstances, non-actions or absences can have a causal impact, such as when my not watering the plants causes them to die. How can an absent factor be a cause for an event if there is no spatiotemporal process connecting them (Henne, Bello, Khemlani, & De Brigard, 2019; S. Khemlani, Wasylyshyn, Briggs, & Bello, 2018)?

Although omissions present a challenge for process theories, psychological studies lend some support for causal process theories. Participant's causal judgements are strongly affected by the presence of a causal mechanism (Lombrozo, 2009; Shultz, 1982). Subjects likewise prefer information on mechanisms over information on co-variation if asked to explain effects (Ahn & Bailenson, 1996; Ahn & Kalish, 2000). Prior knowledge about causal mechanisms that underlie the relationship between a specific cause and the effect it generates is considered to be of importance for causal learning (Ahn & Bailenson, 1996).

1.2.3 Bridging Process and Dependence Theories

Do people judge causation in terms of counterfactual dependence, or more in terms of spatiotemporal processes? While support for both theories can be found, there is no clear evidence for one theory over the other. Some studies demonstrate that causal and counterfactual judgments can come apart (Mandel, 2005; Mandel & Lehman, 1998), while others show that they are closely intertwined (Kominsky & Phillips, 2019; Phillips, Luguri, & Knobe, 2015). Mandel (2005) showed that participants' answers to counterfactual questions about how an outcome could have been avoided can differ from the factors they select as causes for an outcome. In contrast, a study by Phillips and Shaw (2015) directly manipulates counterfactual relevance. They instructed their participants to either reflect on counterfactual alternatives to the causal agent's causal action or, in a control condition, summarise the cover story of the causal scenario. The results demonstrate higher causal ratings for the causal agent when subjects were requested to take counterfactual alternatives to her action into account than in the summary condition.

When directly pitting mechanistic process and counterfactual dependence against each other, Chang and Sanfey (2009) found that participants' causal judgments were most heavily impacted by counterfactual dependence. People did not care too much about whether there was a direct transmission of force between an agent's action and an outcome, i.e. whether an agent actively pushed an object to cause a certain outcome, or whether the agent simply removed an obstacle in the pathway between cause and effect. As long as counterfactual dependence on the action was given, i.e. the outcome was not overdetermined by an additional cause, people disregarded whether there was a direct transmission of force between action and outcome, or not. Ahn and Bailenson (1996), however, show that people's preferred descriptions of causes are based on causal mechanisms rather than co-variation information. Likewise, participants seemed to prefer information on mechanisms over information on co-variation if asked to explain effects. In cases of pre-emption, most people attribute causality to the cause which is physically connected to the outcome (Walsh & Byrne, 2007b).

The conflicting evidence as well as the fact that people's causal judgments are sometimes sensitive to mechanistic processes and at other times solely to co-variation information has led some causal researchers to develop a more unified account (Gerstenberg et al., 2020; Lombrozo, 2010, 2012), and/or to adopt a more pluralistic view towards causation (Godfrey-Smith, 2010; Lombrozo, 2010). Lombrozo (2010) argues that different approaches to causation are used depending on the mode of explanation people engage in. When events are construed teleologically, i.e. with respect to functions and goals, causal judgements are sensitive to counterfactual dependence and relatively insensitive to the presence of physical connections. In contrast, when events are thought of mechanistically, causal ascriptions are sensitive to both counterfactual dependence as well as mechanistic connections

(Lombrozo, 2010; Lombrozo & Vasilyeva, 2017).

Recent developments of counterfactual dependence theories have taken into account the fact that people are sensitive towards causal processes (Gerstenberg et al., 2020; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015a; Woodward, 2002), aiming to dissolve the dichotomy between counterfactual dependence and mechanisms. Building on the notion of difference-making, they acknowledge that a cause can make a difference not only to whether the outcome occurred ("counterfactual necessity"), but also to *how* the outcome came about (Glymour et al., 2010; Lewis, 2004; Schaffer, 2005). How-causation captures the key principle of process accounts, that is, whether there was some kind of transfer of force or energy from the candidate cause to the target (Talmy, 1988; Wolff & Shepard, 2013). Counterfactual contrasts are extended such that these different aspects of causation can be expressed in terms of counterfactual contrasts. In this respect, it is assumed that people not only counterfactually simulate whether the outcome would have been different if the cause had been absent, but also whether the outcome would have occurred differently if something about the causal candidate had changed (e.g. speed, direction of movement, force, etc.) (Gerstenberg et al., 2020; Gerstenberg, Zhou, Smith, & Tenenbaum, 2017). Would the bottle have shattered, or shattered less, if Suzy had positioned her stone slightly differently, or thrown it with lesser force? This way, counterfactual theories can capture whether causal events are connected in a more direct, physical way.

In sum, while process and dependence accounts of causation have often been conceived as rival theories aiming to capture better the nature of how people think about causation, current developments of counterfactual dependence theories aim to bridge this theoretical gap and reconcile both accounts by integrating the aspect of "how"- causation. By construing difference making at a finer level of granularity, rather than just considering what would have happened if the candidate cause had been absent, these accounts also consider what would have happened if the candidate cause had been different in certain physical aspects (Gerstenberg et al., 2020). These aspects can capture mechanism properties that until now have been the mark of causal process theories. The incorporation of the different ways a cause can "make a difference" maintains the counterfactual framework while integrating additional aspects that people deem relevant when reasoning and judging about causation.

1.2.4 Causal Models, Structure and Strength

In the previous section, we have discussed the question what causation is, and how it is represented in people's minds. This section will address the consequent question how people determine what causes there are and how they interact with regards to the effect ("Causal structure"), and in particular, to what extent a causal factors causes the effect ("Causal strength"). Independent of the question of what actually constitutes a causal relation, once the causal connections in a scenario or scene are known, they can be depicted in a causal model. Causal model graphs are a crucial element of causal model theories or, as Bayesian networks, of Bayesian Inference Theories (Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Sloman, 2005; Waldmann, Hagmayer, & Blaisdell, 2006). The qualitative side of a Bayesian network is the graph structure, where a link from C to E corresponds to the claim that E depends on C. Bayesian networks hence consist of nodes that represent variables and arrows that represent the causal relations between these variables. For the *quantitative side*, each variable has conditional probabilities that specify the probability of that variable given the possible values of its immediate causes. The strength of a relationship signifies the degree to or the probability with which the causal variable generates or inhibits the effect variable.

People use a variety of cues such as temporal order, intervention and coherence with prior causal knowledge in order to infer the underlying causal structure. In return, people's causal judgments and responsibility attributions are sensitive to different causal structures (Lagnado, Gerstenberg, & Zultan, 2013). Throughout this thesis, two kind of causal structures will be of particular interest: conjunctive and disjunctive causal structures.

1.2.4.1 Conjunctive Causal Structure

Let us assume two causes C_1 and C_2 for the effect E. In a conjunctive causal structure, C_1 and C_2 are both necessary for E; E is never present if either C_1 or C_2 are absent. However, neither C_1 nor C_2 alone are sufficient to generate E because the effect only occurs if both C_1 and C_2 are present. It is the conjunction of C_1 and C_2 that is brings about E. If, for instance, two people simultaneously stand on an iced lake and as a result break through the ice, it would be a conjunctive causal structure if the ice would not have been broken by either person alone but only by the combination of both people standing on the ice. In the philosophical literature, this structure is often referred to as “joint causation” because the two causal factors jointly cause the outcome.

1.2.4.2 Disjunctive Causal Structure

In a disjunctive case, by contrast, each of the two causes is sufficient to generate the effect, because it suffices if either C_1 or C_2 are present. But neither is necessary because E can be present without C_1, given the presence of C_2, for instance. Consider a case like the one before, but this time, each person would have been enough to lead to the breaking of the ice. The outcome was “overdetermined” since one person would already have brought it about.

1.2.4.3 Causal Strength

A complete causal network, however, does not only contain the qualitative relationships of causal variables represented by the graph, i.e., whether or not there is a causal relation between two events, but also parameters that refer to quantitative aspects. One of these parameters is the *causal strength* of a causal relation (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). The strength of a relationship signifies the degree or the probability with which the causal variable generates the effect variable. Once we have identified C as a causal factor for E, we also want to know to what extent C causes E.

A number of causal strength measures have been discussed in philosophy and psychology. One proposal is aimed at capturing the intuition that a cause C should raise the probability of the effect E above its unconditional value. Another well-established causal measure is causal power that has been introduced within Cheng's power PC theory (Cheng, 1997). Pearl's Necessity-Sufficiency measure is designed to capture the extent to which a cause C is both necessary and sufficient for E (Pearl, 2009). And finally, counterfactual dependence measures take the minimal number of structural changes that have to be made to the situation in order for the outcome to be counterfactually dependent on the cause (Lagnado & Gerstenberg, 2017) (e.g. undo Suzy's action to make the outcome dependent on Billy throwing his stone).

Causal structure and causal strength are the building blocks of our causal knowledge. Using cues like co-variation, temporal order, interventions or prior knowledge, causal reasoners identify causes and effects and find out how they relate to each other. Based on the causal structure, quantitative parameters like causal strength, how likely the effect is given the presence of the cause, can be assessed. While there are various ways to compute causal strength, the aim of this thesis is not to strictly distinguish between these.

In fact, the aim of this thesis is to investigate causal judgments in scenarios in which the causal structure is known, and all causes are necessary and equally strong causes for an effect. If both causal structure and equal causal strength are known, people still show a systematic preference for one cause or the other. In particular, if all causes are equally necessary, many studies show that people select one causal factor in particular as "the cause" of the outcome, or judge one factor as more causal. Conjunctive causal structures have been the focus of interest in most of these studies because they exemplify the phenomenon many philosophers and psychologists struggled with for years: the problem of causal selection (Hesslow, 1988).

1.3 The Problem of Causal Selection: The Role of normality

Consider the example of a forest fire for which the ignition of a cigarette on combustible forest floor and oxygen can be identified as the causes (Hesslow, 1988;

Pearl, 2009). When asked what caused the forest fire, people tend to select one causal factor – the ignited cigarette – and name it as ‘the cause’, whereas the other causal factors are discounted or perceived as mere enabling conditions (Pearl, 2009). Since oxygen and the ignition are equally necessary to generate the fire, the “causal selection problem” arises, that is, the question why people select certain factors as “the” major contributing cause in (conjunctive) causal structures (Hesslow, 1988).

A prominent line of response to this in both philosophy as well as psychology has been to argue that people tend to pick abnormal, rare, or unexpected factors over normal, frequent, and expected ones as causes (Hart & Honoré, 1959/1985; Hitchcock & Knobe, 2009; Kahneman & Miller, 1986). Norms and expectations also play a critical role in how people make causal judgments about omissions, that is, events that didn’t happen. At any given moment in time, many events don’t happen, but we consider only those that violated norms or expectations (Gerstenberg & Stephan, 2021; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Stephan, Willemsen, & Gerstenberg, 2017; Willemsen, 2016). This tendency for abnormal causal factors not only shows in how people make judgements about a *causal structure*, that is what kind of causal factors they pick as causes vs. enabling conditions, but also about *causal strength*. Consider the following case, taken from (Icard, Kominsky, & Knobe, 2017):

> Prof. Smith works at a large university. At this university, in order to get new computers from the university, faculty like Prof. Smith must send an application to two administrative committees, the IT committee and the department budget committee.
>
> The IT committee almost always approves these applications. The department budget committee almost never approves these applications. The budget committee is notorious for turning down almost every application they receive.
>
> Prof. Smith sends in her application. Each committee meets independently and they decide without talking to each other, but their meetings

> are scheduled for the exact same time. The IT committee approves her application, and surprisingly, the department budget committee approves her application. So, Prof. Smith got her new computers.

When asked about the extent to which each of the two committees caused Prof. Smith to get her new computers, Icard et al. (2017) found that participants agreed substantially more with the claim that the department budget committee caused the outcome. Given the unexpectedness of the department budget committee's action, this example demonstrates the impact of the abnormality of an event on people's judgements about "causal strength" (Icard & Knobe, 2016; Icard et al., 2017; Kominsky et al., 2015; Morris, Phillips, Gerstenberg, & Cushman, 2019). In particular, it shows that abnormal objects are not only selected as causes, but this impact occurs with judgments about causal agents as well.

But what exactly is it that makes both the cigarette as well as the budget committee's approval abnormal? Both can be said to violate some sort of statistical or descriptive norm or normality (Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986); that is, both are the less frequent or unlikely factor to happen. Descriptive or statistical norms describe frequencies of entities, behavior or events, indicating whether something is common, habitual or likely (Brennan, Eriksson, Goodin, & Southwood, 2013). An agent's action can deviate from their past behaviour ("*agent-level norm*"), for example when a non-smoker suddenly decides to smoke a cigarette. At the same time, an agent's behaviour can be classified as abnormal with reference to the general behaviour in a group ("*population-level norm*"), for example if an agent smokes while usually being surrounded by non-smokers (Kelley, 1973; Sytsma, Livengood, & Rose, 2012).

Tversky and Kahneman (1974) define an abnormal event or action by the fact that it evokes highly available alternative representations. According to Tversky and Kahneman (1974), an abnormal event often violates expectations and is accompanied by a feeling of surprise. However, it cannot be defined solely by the generation of surprise. A surprising, unanticipated event is not automatically perceived as abnormal, but only if it evokes strong alternative representations. You might be surprised by the high pitch of your colleague's voice, but only judge it as abnormal when you have heard them talking in a low voice before; that is, when there are clear alternative scenarios available. They argue that factors that deviate from a default value are classified as abnormal and therefore more likely to be selected as causes (Kahneman & Miller, 1986)

Why do people judge abnormal causes as "more of a cause"? There are a number of psychological theories that are based on the idea that statistical abnormality or the deviation from the default determines people's causal selection and structure judgments.

1.3.1 Covariation and Correspondence

According to Cheng (1997)'s probabilistic contrast model, factors are selected as causes if they co-vary with the effect within a set of relevant situations reasoners currently consider. A factor is generally considered to co-vary with the effect if the effect is present if the factor is present, and absent if the factor is absent. In consequence, a factor needs to co-vary with the effect in order to be considered causally relevant (Cheng, 1997; Cheng & Novick, 1991). In the focal subset of causal events — ignited cigarette and oxygen —, the cigarette co-varies with the forest fire because the proportion of cases in which there is a fire without an ignited cigarette is much lower than the proportion of cases in which there is a fire and an ignited cigarette. In contrast, oxygen does not co-vary with the fire within this focal set because this factor is constantly present, not only when there is a fire but also when there is no fire. Oxygen is hence considered as an enabling condition (J. L. Mackie, 1974; P. Mackie, 1992), while the cigarette is the cause of the fire.

In a similar manner, the correspondence hypothesis (Gavanski & Wells, 1989; Harinen, 2017) suggests that participants select normal causes for normal effects, and abnormal causes for abnormal effects. Given that it needs both the IT and department budget committee to approve new computers, and given that the department budget rarely approves such an application, the granting of new computers can be considered a statistically rare event. In consequence, people select an unlikely event – the approval by the department committee – as a cause for the outcome that was unlikely to happen.

1.3.2 Counterfactuals, Availability and Sampling

Kahneman and Miller (1986) first proposed norm theory as a theoretical basis to describe the rationale for counterfactual thoughts. Norm theory suggests that the ease of imagining a different outcome determines the counterfactual alternatives created. Abnormal behaviour tends to elicit more counterfactual thought, which elicits stronger affective reactions, for example, the feeling of regret. Exceptional events are also (i.e., taking an unusual route and then getting into an accident) thought to be more mutable than normal events (i.e., taking a usual route and getting into an accident) (Wells & Gavanski, 1989). Counterfactual thought is activated by some negative or abnormal incident, and people's perception of how things go 'normally' determines reference point of what could have gone differently (Roese, 1997; Roese & Olson, 1996).

Building on the long-standing tradition of counterfactual dependence theories of causation (Lewis, 1973), Hitchcock and Knobe (2009) developed an account that construes normality very broadly, including not only statistical norms but also norms of proper functioning and moral or prescriptive norms (Icard et al., 2017; Knobe, 2009; Kominsky & Phillips, 2019; Kominsky et al., 2015; Phillips & Cushman, 2017; Phillips et al., 2015). In a first step, causal relations between causal variables and outcome are identified in terms of counterfactual dependencies as discussed before — the outcome event would not have occurred if the causal variable had been absent. Regarding the bureaucracy vignette, the approval of the IT and budget committee are seen as causal variables because, had either not occurred, the outcome would not have occurred. However, according to this account, people's reasoning process doesn't stop at identifying the causal variables. According to Hitchcock and Knobe (2009), the influence of normative evaluations on causal judgments is attributed to the higher relevance of the counterfactual contrast of the abnormal causal variable. The main idea, similar to norm theory, is that counterfactual reasoning is more relevant for abnormal rather than normal causal factors. The counterfactual alternative to abnormal events is most relevant so that people rather reflect on the alternative that an abnormal causal factor had been absent – and thus normal –, which at least in conjunctive causal structures, leads to a change in the outcome. Intuitively, the alternative to a rare causal event is more relevant than the alternative to a frequent causal event. In consequence, the causal strength of the abnormal causal factor gets highlighted. Thinking about the department committee behaving as usual, that is, the department committee *not* approving the application, involves a counterfactual scenario in which Prof. Smith's application is not granted. This in turn highlights the counterfactual dependence of the outcome on the department budget committee's action, rather than IT committee's action.

Halpern and Hitchcock (2015) formalised the norm-incorporating counterfactual reasoning account of causal judgments. Their idea builds on a causality approach of dependence theories (Halpern & Hitchcock, 2011), introducing the idea that the actual world is compared to several alternative worlds that are ranked by normality. Possible alternative worlds, also called "witnesses" for the factors being causes, are ranked corresponding to their normality. The actual world is more likely to be compared to a normal counterfactual world in which the department committee does not act, rather than to a second, more abnormal world in which the IT committee does not approve. As a result of this normality ordering over possible worlds, people select the budget committee as actual cause (Halpern & Hitchcock, 2015).

Several causal researchers have argued that an event's causal status depends not only on the extent to which the outcome is counterfactually dependent on it, that is, whether it is necessary, but also whether the causal factor is sufficient for the outcome (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020a; McDermott, 2002; Pearl, 2009). Icard et al. (2017) propose a new measure of causal strength that takes into account both necessity and sufficiency, and predicts the influence of normality on causal judgments. According to this measure, causal strength is a weighted function of necessity, and "robust" sufficiency, that is, the extent to which the cause is generally sufficient for bringing about the outcome. The relative weighting of both of these components is determined by the perceived normality of the cause: sufficiency is weighted by normality, while necessity is weighted by abnormality of the cause. Since a cause's necessity and sufficiency are affected differently by the causal structure, this account predicts that the underlying causal structure will affect the influence of normality on causal judgments. In conjunctive structures, the abnormal cause is judged as more causal for the outcome, while in disjunctive structures, however, the normal cause should be judged more causal ("abnormal deflation in disjunctive structures") (Icard et al., 2017). In addition, the account predicts that the impact of normative evaluation concerning the norm-violating cause will also affect causal judgment for the second, norm-conforming cause: In conjunctive causal structures, the causal rating for the norm-conforming cause is supposed to decrease as a result of the other factor's norm violation and the resulting higher relevance of its counterfactual alternative ("Causal supersession") (Icard et al., 2017; Kominsky & Phillips, 2019; Kominsky et al., 2015).

While the accounts presented so far differ with regards to how they explain the impact of norms in causal judgment, they all assume that the normality has a genuine impact on the way people think about causality. Another prominent family of accounts argues that the effect of normality is driven by moral judgments (Alicke & Rose, 2012a; Samland, Josephs, Waldmann, & Rakoczy, 2016a; Sytsma, 2020a). We will come back to this line of theories in more detail in section 1.4.3. Blame-oriented accounts of norms in causal cognition have specifically aimed at explaining the influence of *prescriptive* norms on causal judgments. In fact, what is striking about the influence of normality on causal judgement is that it is not restricted to one type of norms. A large body of research has shown that causal judgments are sensitive to a variety of kinds of norms.

1.3.3 Kinds of Norms

A unique feature of the counterfactual account of norms in causal judgments is that abnormality is used in a very broad sense: not only causal factors that violate descriptive or statistical norms are supposed to stimulate counterfactual reasoning but also factors transgressing prescriptive rules, or functional norms (Alicke & Rose, 2012a; Clarke, Shepherd, Stigall, Waller, & Zarpentine, 2015; Henne, Bello, et al., 2019; Henne, O'Neill, Bello, Khemlani, & De Brigard, 2019; Icard et al., 2017; Knobe, 2009; Kominsky & Phillips, 2019; Willemsen & Kirfel, 2019).

Prescriptive norms such as rules, moral norms or moral obligations have an influence on causal judgments (Alicke, 2000; Knobe & Fraser, 2008). The paradigmatic case for this influence of moral evaluations on causal judgments is the 'pen case' by (Knobe & Fraser, 2008):

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the 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 has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's 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.

In Knobe and Fraser (2008)'s study, participants were asked for their agreement with the two statements "Professor Smith caused the problem" and "The administrative assistant caused the problem". They tended to give higher agreement ratings to Professor Smith who violated a norm than to the administrative assistant who was allowed to take pens (Knobe, 2009; Knobe & Fraser, 2008). The impact of prescriptive norm violations in causal cognition is also demonstrated for causal selection choices by five-year-old children (Samland, Josephs, Waldmann, & Rakoczy, 2016b).

Yet another finding showing the variety of norms is that the violation of norms of proper functioning plays a role for how causal inanimate objects are perceived (Hitchcock & Knobe, 2009; Kominsky & Phillips, 2019). In a variation of the 'pen case', Kominsky and Phillips (2019) employ a scenario in which the academic department has a vending machine, of which one lever gives out pens, and another erasers. When both professor and assistant activate a lever, the one that the professor is using malfunctions and gives out pen, rather than an eraser, causing a lack of pens in the vending machine together with the assistant's activation of the functioning pen-lever. Kominsky and Phillips (2019) find that the artifact that violated a functional norm was judged as having caused the outcome to a greater extent than the properly functioning lever.

What is astonishing about the influence of norms on causal judgment is that it shows up in causal reasoning about both norm-violating causal agents (Knobe, 2009) as well as norm deviating inanimate objects (Kominsky & Phillips, 2019). Since objects and agents belong to distinct ontological categories, the common influence of norms might suggest that we reason about these kinds of causes similarly. In the following section, we want to take a closer look at how people reason about agents and objects as causes.

1.3.4 Agent vs. Object Causation

Philosophers have debated whether object causation and agent causation are the same or distinct kinds of causation (Chisholm, 1976; Lowe, 2001; Pereboom, 2004). Some note that the verb 'to cause' has a different sense when referring to an object than to an action or an agent: our ordinary ways of talking about effects caused by actions support the idea that agent causation is a distinct species of causation (Lowe, 2001; Schwenkler & Sievers, n.d.). But does this linguistic difference map onto a deeper conceptual difference? How do people reason about agent vs. objects as causes?

From early on in cognitive development, humans have particular expectations regarding how psychological states come about and influence behavior, also known as *Theory of Mind* (abbreviated as 'ToM'; we will come back to ToM in more detail in section 1.4). Thus, infants expect social agents – but not inanimate objects – to act in a goal-directed manner (Hamlin, Wynn, & Bloom, 2007; Newman, Keil, Kuhlmeier, & Wynn, 2010). Infants' expectations about the behaviour of physical objects hence contrast with those regarding social agents possessing mental states, and in consequence, affects how they reason about them causally. Newman et al.

(2010) show that children expect agents to causally intervene on the world in a fundamentally different way than inanimate objects do: they infer that an intentional agent rather than a physical force constructed an orderly versus scattered array of blocks. Infants also expect a human hand rather than an artifical tool when they see an improbable versus random sample (regular colour patterns in visual displays) drawn from a population (Ma & Xu, 2013). While non-agentive forces are excpected to increase entropy in the environment, children believe that agents can increase or decrease it at will (W. J. Friedman, 2001).

Psychological studies also show various domain differences in children's reasoning about physical versus psychological events at later stages of development and into adulthood (Baillargeon, 1995). From simple shapes moving around in a two-dimensional space, children and adults can draw complex inferences about the mental states and social interactions of agents (Heider, 1944). The first stage in the process of causal perception is assumed to be the perceptual categorisation of observed entities into either intentional agents or inanimate objects, on the basis of static and dynamic cues. Infants expect that for a physical object to cause another physical object to move, that object must come into direct contact with the other (Leslie & Keeble, 1987), but not in case of social agents. In addition, children understand that physical objects, in contrast to social agents, are not capable of goal-driven self-propelled motion (Saxe, Tenenbaum, & Carey, 2005; Spelke, 1990), leading them to infer a hidden causal agent if the moving object is not be capable of self-generated motion. The fact that children and adults reason differently about psychological and physical ontological domains also influences their expectations regarding the number and types of causes that serve to bring about events in these different domains. Strickland, Silver, and Keil (2017) find that people attribute relatively fewer causes to physical than to psychological events, because physical causal chains are more likely to be construed as simplistic, linear, and deterministic, In contrast, psychological causation is more likely to be thought of as complex and non-deterministic. The tendency to think about psychological and physical events as being embedded in different kinds of causal structures is so pervasive that people reason differently about the causal structure just depending on whether it is framed in psychological versus physical terms (Strickland et al., 2017).

The above evidence suggests a deep divide between the psychological and physical domains for the purposes of causal reasoning. People generally hold inanimate objects and intentional agents in two distinct ontological categories about which they think structurally differently and make different causal inferences. These domain-specific qualitative differences in perceived causal structures may well translate into differing quantitative estimates of the causal strength of the two different kinds of causes. In fact, the causal relationship between agents' mental states and action consequences is typically not perceived as stable as the causal relationship between non-psychologically-driven causes and effects (Juhos, Quelhas, & Byrne, 2015; Walsh & Byrne, 2007a).

In this section, we started out considering the influence of norms and normality on causal strength judgments, and how different types of norms impact people's causal judgments about both agents and inanimate objects. We ended this section, referencing psychological literature that demonstrates some fundamental differences in how people reason about social agents and non-agentive causal factors. Given that people think about social and physical worlds differently, it is not surprising that this also affects how they think about causation in these domains. But what exactly is it that makes us judge the causality of social agents differently than the causality of inanimate objects? In the next section, we will take a look at a factor that has been argued to make (human) living beings unique, and at the same time, plays a crucial factor that people take into account when making causal judgements – our mental states.

1.4 Epistemic States and Theory of Mind

1.4.1 Development of Causal Reasoning

It is easy to take for granted the everyday cognitive abilities that lay the foundation of our social interactions such as our *Theory of Mind*, the ability to attribute mental states to other people and to thereby interpret, explain, and predict their behaviour

(Frith & Frith, 2005; Sterelny, 1990; Wellman, 2011). We explain our own actions by referring to our beliefs, desires, and other mental states, and we attempt to interpret and predict other people's actions by considering their mental states. At the age of two, children begin to be aware that there is a difference between thoughts in the mind and things in the world; they are aware of people's feelings, perceptions, and knowledge (Meltzoff, Gopnik, & Repacholi, 1999; Wellman & Banerjee, 1991) and around the age of three, they use mental-state concepts in their talk (Bartsch & Wellman, 1995). In fact, concepts of mental state emerge rapidly and without formal tuition (Leslie, 1987), and appear to be universal across cultures (Baron-Cohen, 1996; Lee et al., 2018; Sabbagh, Xu, Carlson, Moses, & Lee, 2006). There is more to theory of mind, however, than being aware of mental states. Mental states "mediate" our activity in the world. Two to three-year-old children understand that people act to fulfill their desires (Wellman & Woolley, 1990) and they are able to use information about a person's desire to explain or predict actions or emotions (Bartsch & Wellman, 1995). Infants interpret other people's behavior as goal-directed and assume that these goals are the result of agents trying to maximise subjective rewards they can obtain (Csibra, 2003; Gergely, Nádasdy, Csibra, & Bíró, 1995; Jara-Ettinger, 2019; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Through these assumptions, children draw rich inferences about agents' mental states from their actions and behaviour (Jara-Ettinger, 2019; Jara-Ettinger et al., 2016).

The sensitivity to "agency" and moreover, to agents' mental states, is closely intertwined with their development of causal reasoning. In addition to spatio-temporal properties or statistical relations, whether the agent is perceived as animate or even intentional agent changes whether the agent is is construed as the cause of an event (Muentener & Carey, 2010). Children produce more causal language for intentionally caused events than for unintentionally and object-caused events (Muentener & Lakusta, 2011) and attribute more causality to a human hand engaging in goal-directed action, rather than a mere accidental movement (Leslie, 1984; Muentener & Carey, 2010). Muentener and Carey (2010) find that the deliberateness of the hand movement is necessary to the causal interpretation; if the arm flops down backward behind the screen, after which the box breaks, the infant does not make a causal interpretation, whereas if it arcs forward in a deliberate comparable motion, the infant does does interpret it as causal. While the previous section has already discussed detailed evidence for differences in children's reasoning about agents vs. objects, this research highlights the mental dimension of agency. It is not enough that the potential agent is an intentional agent; in order to interpreted as a cause, they must be also represented as actually acting *intentionally*. Simple cues to intentional agency are sufficient for assigning more causal responsibility to the potential agent. Infants' vast experience with human agents over the first months of their life may facilitate the acquisition and generalisation of human agents' causal properties.

In sum, the close relationship between ToM and the development of causal inference and reasoning suggests a close connection in their execution in fully developed reasoning.

1.4.2 Judgments about Causal Strength

In fact, the sensitivity to agents' mental states in causal reasoning prevails throughout human development. Mental states like intentions, knowledge or ignorance influence causal selection, but also judgments of causal strength, that is *how* causal people perceive an agent for an outcome (Darley & Pittman, 2003; Hilton, McClure, & Moir, 2016; Kirfel & Lagnado, 2020; Lagnado & Channon, 2008; Lombrozo, 2010). For example, the impact of prescriptive norms on causal judgements has been shown to be sensitive towards the agent's mental states. In a widely cited study, Samland and Waldmann (2016) show that people's causal selection judgments were not only sensitive to the abnormality of the causal agent's action, but also to the agent's mental state towards the norm. Agents who violated a norm were chosen less frequently as the cause for an outcome when they did not intend the norm violation or when they did not know that they violated a norm (Samland et al., 2016a; Samland & Waldmann, 2016). Although Samland and Waldmann (2016) used different experiment materials, the pen case from section 1.33. serves as an illustration here. In the studies by Knobe (2009), the rule violating Prof. Smith was judged as more causal than the norm-abiding administrative assistant. The data by Samland and Waldmann (2016) suggest that in a scenario in which the professor had been unaware of the fact that she was violating the norm, she would not be judged as more causal than the normally behaving administrative assistant. Following up on these findings, Kominsky and Phillips (2019) show that epistemic states not only influence causal judgments, but also to what extent people reason counterfactually about the agent's action: people are less prone to think about the norm-violating agent not acting if the agent is unaware of the fact that their action is violating a norm.

In general, people prefer agentive causes over physical causes (McClure, Hilton, & Sutton, 2007; Y. Wu, Muentener, & Schulz, 2016). McClure et al. (2007) find that an agent's action is more likely to be judged causal than a physical object, and that this preference even overrides people's tendency to select proximate causes, that is, even if the human action is in a more distal position and the object in a proximal position to the outcome. However, new studies demonstrate that whether agents are really preferred over inanimate causes depends on the epistemic state they are in (Hilton et al., 2016; Phillips, Young, & Gerstenberg, n.d.). In a series of follow up studies, Hilton et al. (2016) show that people judge actions to be more causally important than a physical object for an outcome only if the agent is aware of the causal opportunity that has presented itself. When varying the length of causal opportunity chains, Engelmann and Waldmann (2021) find that people attribute lesser foreseeability of the outcome to a causal agent the longer the causal chain is. If an action causes harm via a longer causal chain, the harm is seen as less likely and thus less foreseeable than in a direct relation. Likewise, Lagnado and Channon (2008) find that different kinds of foreseeability influence causal judgments. The more the effect of one's action was expected, whether by the causal agent themselves ("subjective foreseeability"), or from an objective viewpoint ("objective foreseeability"), the more causal the agent was judged.

Hilton et al. (2016) investigate the mediating role of foreseeability in judgments about knowing vs. ignorant actions, and test a variety of different hypotheses for the difference in causal judgment about knowing vs. ignorant agents. They find evidence that participants prefer the actions of knowing agents as causes of harm, but not because they are perceived as most controllable. Knowing agents are not privileged as causes over unknowing actions because of perceived (social) controllability. They also showed that the agent's awareness of a causal opportunity did not influence the perceptual organisation of the event sequence, in contrast to the claim that a knowing actions and outcome form a single unit, a "means-end" schema (Hart & Honoré, 1959/1985). They find, however, strong inferences about the deliberateness with which the action brought about the outcome in question and the extent to which the agent desired and believed that the action would bring about that result.

Indeed, acting knowingly has been argued to be closely connected with an intention for the outcome (Malle & Knobe, 1997; Searle, 1980). In "Causation and the Law", legal scholars Hart and Honoré argue for voluntary and deliberate human action as a criterion for selecting legal causes from mere causal conditions, serving as stopping point for causal tracing (Hart & Honoré, 1959/1985). Lombrozo (2010) finds that intentional causal agents that deliberately produce an outcome are rated as more appropriate causes in double prevention scenarios. Even though her account is not focused specifically on *epistemic* states, the account offers a first explanation of why mental states such as intentions might play a fundamental role in causal reasoning. According to Lombrozo (2010), mental states influence the robustness or exportability of the agent cause: An intentional agent is judged as an "insensitive" cause of the intended outcome as the agent will modify their behaviour in response to changing circumstances to bring about the outcome. Romeo and Juliet serve as an illustration here (example from William James' "Principles of Psychology" as cited by Lombrozo, 2010). Unlike iron filings which are stopped from reaching the magnet by an obstacle wall, an intentional Romeo will overcome a variety of obstacles in order to reach Juliet. If Romeo intends to reach Juliet but is blocked by a wall, he will aim to find a way around. Causal robustness is especially sensitive to intentions, since agents tend to realise their plans despite variations in the context and means needed to bring them about. In consequence, intentions increase the exportability of the causal relationship between agent and outcome to different possible scenarios (Heider, 1983; Lombrozo, 2010). Mental states such as intentions mark stable, invariant causal relationships that are exportable to different circumstances (Lombrozo & Carey, 2006; Woodward, 2006).

There is, however, an alternative interpretation of the impact of knowledge states on judgments about causation. Given the crucial role of epistemic states for moral judgments as discussed before, another line of explanation is to account for these findings with reference to moral judgements. Indeed, several theories posit that the influence of agent knowledge and ignorance on people's causal judgements merely reflect their evaluations of responsibility or blame (Alicke, 2000; Alicke, Rose, & Bloom, 2012; Samland & Waldmann, 2016; Sytsma, 2019a), expressed in judgments about causation. While we will come back to these accounts throughout the thesis, we will briefly sketch the three main proponent theories in the next section.

1.4.3 Blame, Responsibility and Accountability

In his *culpable control model of blame*, Alicke (2000) states that the causal contribution of an agent to a negative outcome can be exaggerated if the agent is considered blameworthy. Whether an agent is blameworthy for the outcome of their action is assessed on three different aspects: Volitional behavioural control, volitional outcome control, and causal control. Volitional behavioural control refers to the "mind to behaviour link" and is considered to be present if an agent intentionally performed her action. Volitional outcome outcome control describes the link from "mind to consequence" and is related to the concept of foreseeability, that is, if an agent desired and foresaw the effect event. Finally, causal control can be understood as the strength of the causal connection between an agent's behaviour and the outcome.

While it is assumed that people assess blame based on factoring in all three control components (Alicke & Rose, 2012b), very negative spontaneous evaluations can lead this judgment process into a 'blame validation mode', a kind of processing mode in which the control elements are processed in a way that justifies the spontaneous blame judgment. If people want to blame an agent, for example because the agent intentionally brought about a bad outcome, they mentally exaggerate the strength of the causal connection between the agent's action and the outcome (i.e., the causal control component) in order to justify their initial blame response. It is the because of this exaggerated blame response that people adapt their causal assessment in a biased, post-hoc manner.

Alternative blame-oriented accounts have been developed around the impact of prescriptive norms in causal judgments, but equally apply to the role of epistemic states (Samland & Waldmann, 2015, 2016; Sytsma, 2019a, 2020a). The *accountability hypothesis* claims that the effect of norms or mental states on causal judgments relies on an accountability interpretation of the causal test question (Samland et al., 2016a; Samland & Waldmann, 2014, 2015). This hypothesis is based on the assumption that causal queries are ambiguous. Especially in the context of human actions, Samland and Waldmann (2016) argue that verbal causal test question can both refer to questions concerning the underlying mechanisms of a causal relationship between an action, but also to questions concerning an agent's moral accountability for an outcome. The accountability hypothesis presumes that participants in these studies reflect on the intended meaning of a verbal causal test question and interpret it as a request to assess accountability rather than as a request to assess causality. In particular, socially relevant aspects such as the detection of a moral norm violation and the agent's mental states are important determinants for accountability. The accountability hypothesis assumes that this accountability judgment — rather than a judgment about genuine causation — is reflected in participants' answers to the causal test question about norm-violating or knowing causal agents, expressed in increased causal attributions.

The *responsibility account* (Sytsma, 2020a, 2021) is similar to the accountability hypothesis in some aspects, but goes even further in its claims about how people interpret causal statements. Rather than a pragmatic understanding of the causal test question, this account holds that the dominant use of "cause" in people's ordinary language expresses a genuine normative concept. In this sense, the responsibility account takes the effect of factors such as norms or mental states to be notably broader than the pragmatic account does with regard to context. The responsibility account predicts that people's judgments about causal attribution will *generally* be quite similar to their judgments about normative attributions, and that the normativity of the concept "cause" is not specific to experimental contexts.

1.4.3.1 Interim Conclusion

In sum, while epistemic states do seem to play a fundamental role in the development and execution of causal reasoning, it still remains unclear how and why they do so. In addition, as we have learned in the previous section, the impact of mental states on causal judgment also raises the question whether their influence really targets genuine causal cognition. This thesis aims to resolve the tension between the mounting evidence of the impact of epistemic factors in causal cognition, and the lack of explanation thereof. In consequence, the question of how and why epistemic states influence causal judgments will form a core part of this thesis. In particular, we will address the question of what role agent epistemic states like knowledge or foreseeability play for the impact of norms on causal judgments specifically, but also the question about their more general function in causal cognition and inference.

So far, we have mainly discussed research and theories with regards to the causality of social agents. However, we experience other agents not only as causes in our environment, but also as an important medium for causal knowledge. In fact, our learning process about the world and its causal laws and properties crucially involves other human beings. In particular, we are able to acquire the most sophisticated knowledge — be it causal or non-causal — through other people's explanations. While the introduction so far has mostly focused on causal judgment, the following section will address the process of people's acquisition of causal and other kind of knowledge. We will revisit the various ways of causal learning, with a particular focus on learning from causal explanations. Because explanations usually involve other human beings, both as senders and receivers, they are uniquely interpersonal and often embedded in the context of communicative exchange. Making inferences from what other people say usually also requires making use of Theory of Mind. Given that ToM is involved in this particular interpersonal form of causal learning, is there something special about the way we learn from causal explanations? The last chapter in this thesis aims to find an answer to this question.

1.5 Causal Explanations

1.5.1 Causal Learning

In previous sections, we have already learned that when multiple events are identified as causes for an outcome, normality and causal structure (among other factors) jointly determine which one we select as "the" cause. But how do we reach this causal knowledge in the first place?

In fact, there are multiple routes to causal knowledge. We learn about how the world works through observing what happens, observing others take actions, acting ourselves and learning from the consequences, and by receiving verbal (and non-verbal) explanations (Cheng, 1997; Cheng & Novick, 1991; Lagnado et al., 2007; Waldmann & Hagmayer, 2001). Infants can infer the causal structure of the world by learning novel cause-effect relations from observing other people's actions. (Gergely, Bekkering, & Király, 2002; Meltzoff, 1988; Meltzoff, Waismeyer, & Gopnik, 2012; Waismeyer, Meltzoff, & Gopnik, 2015). Observation and imitation are a helpful causal learning mechanism for children because adults and older siblings engage in a wider range of causal interactions than they can do themselves (Meltzoff et al., 2012). However, causal links often are not directly observable. In these cases, we must infer causal relations from the observable data points that are available to us. Several cues have been shown to guide the process of translating covariational data into representations of cause and effect, for example temporal order (Lagnado & Sloman, 2006; Lagnado et al., 2007), spatial and temporal contiguity (Shanks, 1989), contingency (Kushnir & Gopnik, 2007; Msetfi, Wade, & Murphy, 2013) as well as outcome density (Dickson & Theobald, 2011; Vallée-Tourangeau, Murphy, & Baker, 2005) When different possible causal structures are compatible with an observation, interventions enable us to differentiate between them (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; D. Sobel & Sommerville, 2010).

Finally, people also rely on other people's explanations and instructions to gain causal understanding about the world. The ability to *explain* is at the core of how we understand the world and ourselves (Craik, 1943; Salmon, 1984; Woodward, 2003). As scientists, we are often not content with merely being able to predict what will happen. Instead, we strive for a deeper understanding of the underlying causal laws or mechanisms that dictate *how* and *why* the world works the way it does. Explanations also play a critical role in our everyday lives (Hagmayer & Osman, 2012; Heider, 1958; Lombrozo, 2006). One of the most important functions of causal explanation is in interpersonal interaction: we understand one another as guided by reasons, goals and desires, with much of behaviour intelligible in decidedly causal terms (A. R. Buss, 1978; Davidson, 1963; Malle, 1999). In the following, we want take a closer look at causal explanations, and the ways we learn from them.

1.5.2 Types of Causal Explanations

Although very few explanations contain an explicit causal statement of the form "C caused E", most explanations involve some kind of causal consideration. Explanations and causes are considered to be closely linked, with many instances of causal judgments having explanatory value or featuring in explanatory considerations, and in return, many explanations referring to causes or causal relations (Keil, 2006; Lombrozo, 2012; Lombrozo & Vasilyeva, 2017). It is common to distinguish between four different types of explanation: *causal explanations*, referring to causal relations, *mechanistic explanations*, referring to material processes and components, *teleological explanations*, referring to goals or functions, and finally *formal explanations*, citing properties, type, or category membership (see also Lombrozo, 2012) (Keil, 2006; Lombrozo, 2012; Lombrozo & Vasilyeva, 2017).

Studies show that all these explanation types can reveal causal concepts in some way or another, and prompt different ways to think about the underlying causal structure (Lombrozo, 2010; Lombrozo & Vasilyeva, 2017). For example, mechanistic explanations have been shown to highlight the transmission of momentum or energy between a cause and an effect (Lombrozo, 2010). In a similar manner, functional explanations are only accepted when the stated function of the explanandum also reveals something about the causal process by which it came about ("The woodpecker's beak enables it to catch insects under the bark of trees [and therefore, to survive evolution]") (Lombrozo, 2006; Wright, 1976). In contrast, it is less clear how formal properties convey information about causal relationships. According to Prasada and Dillingham (2006, 2009), formal explanations in the form of token/type categorisations ("Fido is a dog") define a constitutive, but not necessarily causal relationship. Lombrozo and Vasilyeva (2017) however suggest that some formal explanations can be understood causally by pointing to the category-associated essence that is causally responsible for the property that is explained ("Fido is barking because he is a dog.") (Cimpian & Salomon, 2014; Lombrozo & Vasilyeva, 2017). In sum, while there is a variety of explanations, they all enable causal learning in some way or another (Lombrozo, 2010; Lombrozo & Vasilyeva, 2017).

Despite the prominence of explanation throughout human affairs, we still lack a detailed understanding of *how* inference from explanation actually works. As many have emphasised (M. Friedman, 1974; Keil, 2006; Lombrozo, 2006; van Fraassen, 1980), explanations – conceived as answers to "why?" questions – facilitate *understanding* on the part of the individual receiving the explanation. This observation highlights a significant *communicative* dimension of explanation. While theorists have often attempted to study the subject in relative abstraction from concrete discursive contexts (e.g., Lewis, 1986; Salmon, 1984; Strevens, 2008), a number of researchers have argued that many idiosyncratic features of explanation demand that we take this communicative dimension more seriously (Hilton, 1990; Potochnik, 2016; Turnbull & Slugoski, 1988). In this light the main question becomes: How exactly do we employ causal explanations to impart understanding? That is, what kinds of strategies do we use to produce and interpret causal explanations in communication?

1.5.3 Learning (more) from Causal Explanations

The particular communicative dimension of causal explanations prompts the question whether there is something about them that separates causal learning by explanation from causal learning by observation or intervention. Take the following example: Your friend Suzy is trying to get her paper accepted, and she needs to convince two reviewers in order to achieve this. When discovering her article in your Google scholar suggestions, you learn that she finally got her paper accepted, and hence must have convinced both reviewers. However, in contrast, imagine you learn about her acceptance through her explanation "The paper was accepted because I managed to convince Reviewer 2." Does it make a difference if we learned about Suzy's paper acceptance by her explanation, rather than by inference from observation? From your friend's explanation, you learn that her paper got accepted, and hence that she must have been able to convince both reviewers. However, do we learn something more in the case of learning by explanation? For example, that it must have been particularly difficult to convince Reviewer 2?

While numerous studies in the psychology of inference have researched how people infer cause-effect relationships from observing and interacting with the world (Cheng, 1997; Cheng & Novick, 1990; Gopnik et al., 2004; Lagnado et al., 2007; Waldmann & Hagmayer, 2001), to date no one has investigated whether people infer more from causal explanations than just a causal connection. Studies show that the goodness of an explanation is determined not only by whether it's accurate, but also by being a relevant and felicitous contribution to a conversation and social exchange (Hilton, 1990). What kind of causal explanation is considered to be relevant for a conversation is thereby largely dependent on the inquirer's mental model of the target event, and the gaps in their knowledge (Hilton, 1996; Slugoski, Lalljee, Lamb, & Ginsburg, 1993). However, little is known about the kind of inferences we draw from causal explanations, especially with regards to their communicative function. That is especially true for causal explanations of particular events, that is, explanations of the kind "Germany lost in the UEFA Euro 2020 because of England." Although these kinds of – short and concise – explanations are a regular and important mode in communicating and learning about the world, it is not entirely clear how much exactly we can learn from them. Do we merely learn that the English team is a cause of Germany's exit from the European Football Championship? Or are we able to infer more, for example, that the English team was surprisingly strong this year? When confronted with an explanation, it seems we can rely on a somewhat shared understanding between us as listeners and the speaker, such as the tendency to explain via reference to unexpected, abnormal causes (Hilton & Jaspars, 1987; Kahneman & Miller, 1986). It is this tacit assumption that allows us to deduce rich information from short explanations. This inference however draws on the underlying ability to simulate the speaker's preferences or thoughts at the time of communicating the explanation. It requires some basic theory of mind.

1.5.3.1 Explanations and Theory of Mind

In fact, the ability to attribute mental states to oneself, and to others is central to providing an explanation to another agent, but also for making inferences from another person's explanations. Let's consider the following example, taken from Shvo et al. (2020):

> Mary, Bob and Tom are housemates sharing a house. While Tom was away on a business trip, Mary and Bob noticed a hole in the roof of their house and called a handyman to fix it. Before the handyman could come, however, it rained during the night and the floor got wet. Bob, who sleeps in a windowless room, did not notice the rain. Tom, who just got back from his trip that day, noticed the rain but did not know about the hole in the roof. Mary saw Tom return to the house at night and so knew that Tom knew that it had rained. In the morning, when trying to explain the wet floor to Bob, Mary tells him that it had rained during the night and when explaining to Tom she tells him that she and Bob had discovered a hole in the roof (adding that the handyman will arrive the next day).

The scenario illustrates how Mary tailors her explanations to each of her housemates. The information she is providing to both of them is sufficient to explain the wet floor with regards to their respective epistemic states, that is, what they as listeners know about the situation. In order to give adequate explanations, Mary makes use of her ability for theory of mind — her ability to attribute mental states to herself and to others. Indeed, research shows that people flexibly adapt an explanation when their counterpart is ignorant or shares an only partly overlapping view of the explained event (Slugoski et al., 1993). In a study by Slugoski et al. (1993), participants gave different explanations to different explainees (i.e., the recipient of an explanation), based on their beliefs about the beliefs of the explainees. Explainers may also omit a factor in an explanation if they consider it irrelevant or redundant to the question they have been asked (Einhorn & Hogarth, 1986; Hilton & Jaspars, 1987). Hence, explanations are not only constructed in a way that leads to the best understanding on the listener's side; whether a statement is perceived as an explanation in the first place depends on whether it generates understanding in the recipient (Waskan, Harmon, Horne, Spino, & Clevenger, 2014; Wilkenfeld & Lombrozo, 2018). An explainers' explanation is only as good as their ability to model the mental states of the people that receive the explanation, and how their knowledge will change in light of the explanation.

In sum, while we there are varieties of ways we can learn about causality, causal learning from explanation presents a special case. The study of learning from causal explanations has often neglected their significant role in human communication. Embedded in communication, explanations serve and facilitate the coordination between speaker and listener – despite, or rather because as listeners, we sometimes lack relevant information about the world. In the last chapter of this thesis, we will investigate the question of what people are able to infer from causal explanations, capitalising on their communicative, interpersonal function. Yet again, we will assume that our theory of mind abilities play a crucial role in the process of drawing inferences from explanations.

1.6 Road Map of this Thesis

In what follows, we will investigate the role of epistemic states and theory of mind in causal judgments. All three research projects in this thesis tackle this question, from different angles, some more directly, and some rather indirectly. Overall, this thesis argues that we need to think more closely about the relation between causal cognition and theory of mind.

The first chapter ("Abnormal Agents and Epistemic States") will return to the 'problem of causal selection, as described and discussed in the introduction. In this project, we aim to take a different angle on this question. We ask whether considering causal agent's epistemic states, and in particular how these change given the normality or abnormality of their actions, might help explain why normality influences causal judgments.

In the second chapter ("Epistemic Interventions in Causal Reasoning"), we ask more generally why it is that the epistemic state of a causal agent plays a role for how causally we judge their actions. Drawing on the rich tradition of counterfactual dependence theories of causal judgments, we aim to reconcile the influence of epistemic states with this theoretical account. We put forward the hypothesis that epistemic states function as point of interventions in people's counterfactual reasoning. In four experiments, we find evidence that agents' epistemic states and epistemic conditions not only influence causal judgments about agents, but also the counterfactuals that people consider.

In the third and final chapter ("Inferences from Explanations"), we turn around the general research programme in causal cognition research, that is, identifying factors that influence people's causal judgements and causal explanations. More precisely, we aim to investigate the kind of inferences that people draw from causal explanations. Once again, we draw on the vast amount of research demonstrating the impact of norms on causal judgements, but this time, as a factor that people infer from those judgments. In two sets of experiments, we show how people are able to infer normative and causal information from causal explanations that go beyond what is explicitly communicated. Adopting a communication-theoretic account of causal explanations that relies on interlocutors' ability for theory of mind, we explain how this can be the case.

In the conclusion, we reflect on the general relationship between theory of mind and causal cognition.

Chapter 2

Abnormal Agents and Epistemic States

"*The person who lit the match ought to have anticipated the presence of oxygen, whereas nobody is generally expected to pump all the oxygen out of the house in anticipation of a match-striking ceremony.*" (Pearl & Mackenzie, 2018, "The Book of Why", p. 291)

2.1 Introduction

Dust explosions, caused by the presence of combustible dust particles and a source of ignition, present a real danger in a variety of workplaces and industries. In 1983, damage of $1.2 million was caused by a fire in a furniture manufacturer in Walkertown, North Carolina. An employee who was carrying out repair works on the rooftop of the factory had dropped a cigarette in wood dust accumulated from the ventilation systems on the roof. Although both the wood dust and the lit cigarette were necessary for the fire to occur, the employee's smoking was reported as the main cause of the fire (U.S. Chemical Safety and Hazard Investigation Board, 2006). However, when a dust explosion occurs, the source of ignition is not automatically determined as the major cause. At an outdoor music festival in a Taiwanese fun-park in 2015, a festival-like colour powder was released over an area where party-goers were dancing and smoking, eventually leading to a dust explosion. The local fire department reported the spray of the combustible colour powder to have caused the incident ("Taiwan Formosa Water Park explosion injures hundreds", 2015, June 28).

Accident and incident reports give a unique insight into our judgments about actual causation. Which factors are determined to be crucial for an event and which are deemed peripheral shows how we attribute causality among a set of multiple causal factors. Why is the lighting of a cigarette determined as "the cause" for the dust explosion in the furniture factory, but not for the explosion in the water park? A prominent pattern found in causal cognition research is people's tendency to single out atypical or abnormal events as causes (Hart & Honoré, 1959/1985). And in fact, while it is fairly common to smoke at an outdoor festival, spreading combustible dust is rather atypical. In contrast, woodworking is a typical activity in a wood factory, while smoking, even in designated places, is comparatively rare. People's preference for atypical actions, objects or events as causes is a well-studied phenomenon in philosophy and psychology (Cheng & Novick, 1991; Hart & Honoré, 1959/1985; Hesslow, 1988; Hilton & Slugoski, 1986; J. L. Mackie, 1974), and might well explain why causal selection in the two dust explosion cases differs.

2.2 Aim of this chapter

In this chapter, we aim to provide an alternative account of people's increased causal attributions to 'abnormally' acting agents ("abnormal inflation" effect, Icard et al. (2017)). In addition to acting atypically, both the smoking employee as well as the event organiser who instructed the colour powder release could have anticipated that the other would have acted 'typically'. In consequence, both could have foreseen the outcome of their action. We argue that it is the *epistemic state* of an abnormally acting agent, – what the agent foresees or expects – rather than the mere abnormality of their action, that is driving people's causal judgments in such cases. In four experiments, we show that the difference in causal judgements about atypical and typical actions is driven by the difference in agents' epistemic states. Furthermore, this epistemic asymmetry influences people's inferences about the agents' mental states towards the outcome. These findings shed light on the crucial role of agentive epistemic states for causal attributions, and their role in people's preferences for atypical causes.

2.2.1 Abnormal Agents

How people perceive and judge atypical actions and events, often also termed "exceptional" (Kahneman & Miller, 1986) or (statistically) "abnormal" (Icard et al., 2017; Knobe, 2009), is of central interest to psychologists and philosophers. In particular, the atypicality an action has been shown to influence a variety of properties that people assign to the acting agent. In an experimental scenario by Kahneman and Miller (1986), Mr. Jones decides to give a stranger a ride and gets robbed as a consequence. People were more likely to attribute to Mr. Jones a feeling of regret over his action when Mr. Jones usually would not take hitch-hikers in his car, compared to when he frequently did so (Kahneman & Miller, 1986; Kutscher & Feldman, 2019; D. T. Miller, Turnbull, & McFarland, 1990; Roese, 1997). Similarly, in the case of negative personal consequences, an abnormally acting agent is perceived as more unlucky than an agent who acted as usual (Kutscher & Feldman, 2019), and as more deserving of compensation (Kutscher & Feldman, 2019; D. T. Miller & McFarland, 1986, see the former for a failure of replication). At the same time, people attribute more free will and free choice to an agent who deviates from a usual routine, and judge them more responsible for a harmful outcome (Fillon, Lantian, Feldman, & N'gbala, 2019; D. T. Miller & McFarland, 1986).

Of the various cognitive domains that have been shown to be influenced by normality, causal cognition is perhaps the most surprising one. Among a set of multiple causes, people systematically select the factor that is most abnormal as "the cause" for an outcome (Cheng & Novick, 1991; Hart & Honoré, 1959/1985), or rate this factor as having caused the outcome to a greater extent (Icard et al., 2017; Knobe, 2009; Kominsky, Phillips, Knobe, Gerstenberg, & Lagnado, 2014). Crucially, this causal preference prevails even when all factors are necessary for the outcome to occur, exemplified by the common practice to cite the lit match as cause of a fire, but neglect to mention the equally necessary oxygen in the air (Pearl, 2009). Consider the following example, adapted from Icard et al. (2017):

> A designer and a travel agent work in the same building. The build-

> ing's climate control system is a new design that saves energy by keeping track of the number of people in the building, and only turns on when the designer AND the travel agent enter the building. The travel agent almost always arrives at work at 8:45am, but the designer almost always arrives at 10am. One day, the travel agent arrives at 8.45am, and, unexpectedly, the designer also arrives at 8.45am. As a result, the climate system turned on at 8:45am.
>
> (Icard et al., 2017, Vignette 'Building')

When asked about the extent to which each of the two people caused the climate control system to turn on at 8:45, Icard et al. (2017) found that participants agreed substantially more with the claim that the designer, the atypically acting agent, caused the outcome. Their study demonstrates how, despite equal causal contribution, the frequency or "normality" with which an agent performs an action influences how causal they are perceived for an outcome (Icard et al., 2017).

2.2.2 Normality in Causal Cognition

As outlined in the introduction to this thesis, people's tendency to attribute increased causality to abnormal rather than normal causes has been shown for atypical or unexpected events ("statistical" or "descriptive normality"), but also for events that violate social or moral norms ("prescriptive normality") (Henne, Niemi, et al., 2019; Icard et al., 2017; Kirfel & Lagnado, 2018; Knobe, 2009; Kominsky et al., 2015). The abnormality of a cause affects judgments about its causal strength, but also about the causal strength of other causal co-factors. If two causes jointly cause an outcome, people increase causal attribution to the abnormal causal factor, and at the same time reduce how causal they judge the other, normal causal factor ("causal supersession", Kominsky et al. (2015)). Norms or normality have also been shown to play a crucial role for causal judgements about omissions, i.e. events that did *not* occur (Henne, Niemi, et al., 2019; Henne, Pinillos, & De Brigard, 2017; McGrath, 2005; Sartorio, 2009; Willemsen, 2018; Willemsen & Reuter, 2016). From a wide range of events that did not occur, people select those as causes that violated norms or prior expectations (Henne, Niemi, et al., 2019; Willemsen, 2018; Willemsen & Reuter, 2016) Similarly, the common tendency to pick most recent events as “the cause” for an outcome is overridden if prior events in a chain of causes are perceived as more abnormal (Reuter, Kirfel, Van Riel, & Barlassina, 2014).

As outlined in the introduction, there is a variety of competing accounts aiming to address people’s preference for abnormal causes. Normality has been suggested to influence people’s mental representation of alternative possibilities, or *counterfactual reasoning* (Epstude & Roese, 2008; Gerstenberg & Icard, 2020; Henne, Kulesza, Perez, & Houcek, 2020; Icard et al., 2017; Kahneman & Miller, 1986; Knobe, 2009; Roese & Epstude, 2017). According to counterfactual accounts, people show a general preference to mentally undo an abnormal action, and generate more alternatives to an agent deviating from typical behaviour. As a result, more counterfactual alternatives come to mind in which the outcome would have been absent as a result of a change in the abnormal agent’s behaviour, emphasising this agent’s perceived causal strength.

Another family of accounts argues that these effects are driven by moral judgments (Alicke & Rose, 2012a). On one account, prescriptive norms have been claimed to influence causal judgments by normative evaluations such as the attribution of blame or responsibility (Alicke & Rose, 2012a; Driver, 2008; Sytsma et al., 2012). Alternatively, the presence of norms in these kind of experimental scenarios has been suggested to shift participants’ interpretation of the causal test question about agents into the realm of accountability (Samland & Waldmann, 2016).

Additional theories have been suggested with reference to co-variation of cause and effect (Cheng & Novick, 1991; Harinen, 2017) or general pragmatic principles of communication (P. Grice, 1989; Hilton & Jaspars, 1987; Kirfel, Icard, & Gerstenberg, 2020). In sum, there is general consensus that deviations from normality influence people’s causal attributions, yet there is an ongoing debate about why these effects persist.

2.2.3 Revisiting Atypicality

The dust explosion cases as well the "building" vignette demonstrate our tendency to select abnormal or atypical actions as causes. As indicated earlier, however, upon closer examination it becomes clear that the agents differ in yet another aspect. Let's consider again the 'Building' vignette from (Icard et al., 2017). In comparison to the travel agent who presumably did not anticipate the designer's earlier arrival, the designer can expect the travel agent will arrive at 8.45am. Assuming that employees are familiar with the climate control set up, by deciding to arrive at 8.45am, the designer hence is likely to foresee the the climate control turning on, or more so than the travel agent. This basic epistemic difference between the agents might give rise to further inferences from abnormal behaviour: Did the designer want the climate control to turn on that day? Did they intend to turn it on?

The influence of abnormality on causal judgments about agents has been predominantly shown for causal structures in which two causes are necessary for the occurrence of the outcome (Icard et al., 2017; Knobe, 2009; Kominsky et al., 2015; Sytsma et al., 2012). In such cases of "conjunctive causation", the causal consequences of one agent's action is dependent on the action of another agent. This causal co-dependence also maps onto the agent's knowledge about the consequences of their action. Foreseeing the consequences of one's own action is to some extent dependent on knowing what the other agent does. The frequency of past behaviour is one of many social cues that are used for predicting the behaviour of others (Danner, Aarts, & de Vries, 2008; Epstein, 1979): How 'normal' agents act hence influences the predictability of their actions (Rivis & Sheeran, 2003). This relationship between normality and predictability of actions applies to various kinds of normality or typicality that have been discussed in the literature: *token - or agent - level typical behaviour* ("Ben usually smokes at partys"), *type - or group - typical behaviour* ("Ben's friends usually smoke at parties."), but also the *typicality of features or properties* ("Wood dust is usually combustible."). In sum, for two agents (groups) in a conjunctive causal structure whose actions differ in any of these kinds of typicality, it follows that they will also differ in the extent of knowing what the other agent does, and therefore in knowing the consequences of their actions.

2.2.4 To Know or Not to Know: The Role of Epistemic States

Epistemic states, what an agent thinks or believes, and mental states more broadly, what an agents wants, feels or desires, play a crucial role for moral judgements, but also influence attributions of causation more directly (Kinderman, Dunbar, & Bentall, 1998; Lagnado & Channon, 2008; Sytsma, 2019a). While we have discussed some of this evidence in the introduction already, a quick summary might serve as a reminder of the vast impact of mental factors on judgments of causation. Intentional actions are rated as more causal than unintentional actions with accidental outcomes (Alicke et al., 2012; Gilbert, Tenney, Holland, & Spellman, 2015; Wiener & Pritchard, 1994; Williams & Lombrozo, 2010). Lagnado and Channon (2008) tested causal and blame attributions to agents in causal chains and found that both intentionality and foreseeabilty increase these attributions to the agent. People judge an agent causing a foreseeable outcome, both from the agent's perspective ("subjective foreseeability") as well as from an objective point of view ("objective foreseeability"), as more causal than if the outcome was unforeseeable.

Epistemic states influence causal judgements, and recent studies suggests they also mediate the effect of normality on causal judgments. The influence of prescriptive norm violation on causal judgments has been shown to hinge largely on the knowledge state of the norm-violating agent (Samland et al., 2016a; Samland & Waldmann, 2015, 2016). An agent who unknowingly performs a forbidden action, e.g. by being unaware about the rule or norm that they are violating, is not judged more causal than someone who abides by the norm. Sytsma et al. (2012) show that typical, rather than atypical behaviour, is judged more causal if repeated behaviour increases the agent's ability to foresee or anticipate an outcome (Kirfel & Lagnado, 2017; Sytsma, 2020a; Sytsma et al., 2012).

However, epistemic states not only mediate causal attributions to (atypical) behaviour; there is evidence that deviations from normal behaviour trigger inferences about a wider class of mental states of the agent (Alicke, 2000; Gerstenberg et al., 2018a; Knobe, 2003; Monroe & Ysidron, 2021; Sytsma, 2019a). Research in attribution theory has traditionally argued that unexpected or odd behaviour is diagnostic of that agent having certain mental states, or dispositional and internal attributes (Jones, Davis, & Gergen, 1961; Jones & Harris, 1967; Kelley, 1967, 1973; Lucas, Griffiths, Xu, & Fawcett, 2009; Uttich & Lombrozo, 2010). People draw strong inferences about an individual's motives, intentions or character when their present action deviates from past behaviour (Heider & Simmel, 1944) or general expectations (Jones et al., 1961; Jones & Harris, 1967). Engaging in a prescriptively or statistically 'abnormal' behaviour receives higher attributions of free will (C. J. Clark, Baumeister, & Ditto, 2017; C. J. Clark et al., 2014), mediated by an inference about the agent's particularly strong personal desire for and choice of the abnormal action (Monroe & Ysidron, 2021).

Despite the often crucial role of epistemic states in causal attributions and inferences from abnormal behaviour, they are rarely controlled for experimentally. Studies on the effects of normality on causal judgments have predominantly used descriptive vignettes with human causal agents (but see Gerstenberg & Icard, 2020; Kirfel et al., 2020; Kirfel & Lagnado, 2019). The short verbal description of these causal scenarios often lack in-depth information about what the causal agents think, believe or know. As Sytsma (Sytsma, 2019a, p. 25) points out: "This [the lack of control for mental states] raises an important methodological issue for empirical work on ordinary causal attributions: researchers need to carefully consider and control for the inferences that participants might draw concerning the agents' mental states and motivations".

2.2.5 Hypotheses

In this chapter, we aim to investigate what role agents' epistemic states play for causal judgments when the statistical normality of agents' actions varies. Previous research has explained the difference in causal judgements about an abnormal and normal agent in a conjunctive causal structure with reference to the difference in normality of behaviour (Hitchcock & Knobe, 2009; Icard et al., 2017; Knobe, 2009; Kominsky et al., 2015). Here, we explore the question whether the co-occurring epistemic asymmetry between abnormal and normal agent is the factor that influences people's causal attributions. In addition, we also aim to investigate what inferences people make about the causal agents' mental states from the normality of their behaviour.

2.2.5.1 Causal Judgements

Our main hypothesis derives from the close connection between normality and predictability of behaviour (Danner et al., 2008). In a conjunctive causal structure, expectations about the other agent's actions influence the relative foreseeability of the consequences of one's own action. An agent whose co-agent acts typically will hence be in a better position to foresee whether their action will cause the outcome, compared to someone whose co-agent acts atypically. Comparing a normal and abnormal agent, the latter is provided with an epistemic advantage. At the most basic level, our hypothesis is that in a conjunctive causal structure, it is the abnormal agent's foreseeability of the outcome (via the foreseeability of their normal co-agent's actions) that leads to an increase in people's causal contributions to the abnormal agent.[1] Crucially, however, whether this epistemic advantage arises is dependent on whether the two agents know about each other and each others' action frequency. If agents do not know about how often the other one acts, no asymmetry in forseeability of the outcome arises. Our hypotheses with regards to causal judgments about intentional agents in a conjunctive causal structure are as follows:

(i) When the agents know about each other's actions, people will judge the abnormal agent as more causal for the outcome than the normal agent (*epistemic advantage*).

[1] The example of the dust explosion cases suggests that the influence of epistemic states might go beyond occurrently entertained beliefs and expectations (Zimmerman, 1997). Even if party organiser and smoker did not currently or consciously expected their action to have a certain consequence, they could or should have (reasonably) been expected to do so (or more so than others). In this sense, non-actual *dispositional* epistemic states (FitzPatrick, 2017; S. Murray & Vargas, 2018; Sher, 2009), and perhaps even normative epistemic states (FitzPatrick, 2008) might have an equal impact on causal judgments. In this chapter, our aim is to establish the influence of epistemic states in causal judgments about abnormal causal agents. As a test case, we will use actual and prevailing beliefs and expectations. We leave open the possibility of this assumed influence to expand to weaker or normative modes of epistemic states as well.

(ii) When the agents have no, or limited knowledge about each other's actions, people will not judge the abnormal agent as more causal for the outcome the normal agent (*no epistemic advantage*).

2.2.5.2 Outcome-Oriented Mental State Inference

In the last part of the chapter, we return to the idea that the normality of behaviour not only influences causal judgments, but also gives rise to further inferences about the agents' mental states (Jones & Harris, 1967). We develop a Bayesian network model that allows us to predict the probability of an agent's mental state toward an outcome based on the normality of the agents' behaviour, as well as their epistemic states. With the help of a simplified example, we illustrate the model's prediction for a case in which the agents differ in the frequency of their actions and know/don't know about each other. We then test the qualitative predictions of the model for a case in which the agents know about each other, and a case in which the agents do not know about each other. In line with our previous hypotheses, we predict that people will infer outcome-oriented mental states to a greater degree from abnormal behaviour, but only when the abnormal agent has an epistemic advantage:

(iii) People infer outcome-oriented mental states to a greater degree from abnormal behaviour than from normal behaviour, but only if the agents know about each other.

We conducted four experiments to test these hypotheses. Experiment 1 will test hypothesis i). Experiment 2 will test hypothesis ii). Experiment 3 will test both hypotheses i) and ii) for a case in which the agent's expectations about each other generalise to a novel context. In Experiment 4, we put hypothesis iii) to test by assessing people's inferences about outcome-oriented mental states of agents who acted normally or abnormally. In sum, our experiments explore the influence of epistemic states for causal attributions to and inferences from abnormal behaviour.

While the hypothesis we put forward here is neutral with regards to the exact mechanism by which epistemic states influence causal judgements, we will address this point by returning to some of the accounts of normality in causal cognition in the General Discussion.

2.3 Experiment 1: Abnormality and Knowledge about Each Other

In the first experiment, we aimed to replicate previous research on the influence of atypicality on causal judgements. For this, we followed the general experimental paradigm involving two intentional agents in a conjunctive causal structure. The study was designed to explicitly control for participants' assumptions about agents' epistemic states. In Experiment 1, we aimed to test the effect of atypicality when both agents clearly know about each other and each other's actions.

2.3.1 Participants and Design

We recruited 159 participants via Amazon Mechanical Turk (80 participants per condition). Five participants were excluded for failing half of comprehension questions, i.e. five or more of the ten comprehension check questions in the entire study (see Appendix A), leaving a final sample size of $N = 154$ (M_{age} = 33.28, SD_{age} = 9.72, N_{female} = 54). The experiment has a 2 normality ("Both agents normal" vs. "Agent 1 normal, Agent 2 abnormal") × 2 agent (Agent 1: fixed agent vs. Agent 2: varied agent) × 2 scenario ("kettle" vs. "microwave") mixed design. The factors 'normality' and 'agent' were manipulated within participants, 'scenario' was manipulated between participants. All materials, data analyses, model code and power analyses can be found at `https://github.com/LaraKirfel/Atypicality`.

2.3.2 Material and Procedure

Participants completed both experimental conditions, i.e. the *'Both agents normal'* condition, and the *'Agent 1 normal, Agent 2 abnormal'* condition. It was randomised which condition participants completed first. Each of the two experimental

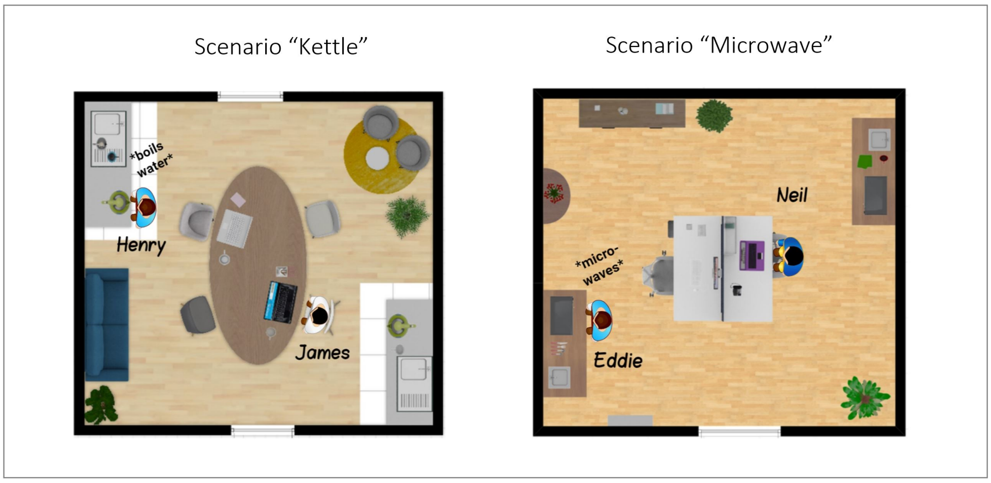


Figure 2.1: Illustration of two scenes from the video clips in Experiment. In each scenario, two agents work together in a joint office. Depending on the scenario condition, the office has two kettles, or two microwaves that the agents can use. In the scenario pictured, only one agent uses the item ('Agent 1 normal, Agent 2 abnormal' condition).

conditions was structured in a similar manner. First, participants received a short introductory text about the scenario, together with a picture of the scene. The scenario involved two co-workers who share an office together (Figure 2.1). The agents were male with different names and looks across all conditions. Depending on the scenario type, the office either has two kettles or two microwaves that the employees can use whenever they want. For energy saving purposes, the company introduces the "Green Friday" on which the building is switched into a power-saving mode. As a result of this power-saving mode, the use of *more than one* kettle [microwave] in the office on Fridays will lead to a power cutout in the company, and agents are aware of this policy. This power-saving mode was introduced as a purely causal mechanism, with no particular prohibition or ban of using these items. Any inference about this mechanism as suggestive for a norm would still apply equally to all agents. The introduction text also stated explicitly that the agents share an office, know each other very well, and usually know what the other is doing during the day. The introduction was followed by four comprehension questions, asking about the office situation (1), the agents' knowledge about each other (2), and the underlying causal structure (1) (see Appendix A).

2.3.2.1 Causal Structure

Studies on the role of norms in causal cognition have predominantly used vignette studies. In these vignettes, the described time frame of the causal event focuses on the narrow time point at which the singular causal outcome occurs. Prior causal history, e.g. whether the outcome has occurred before, is ambiguous. The co-variation between cause and effect has been shown to influence judgments of causal strength (Cheng & Novick, 1991; Harinen, 2017; Kirfel & Lagnado, 2018; Lagnado et al., 2007). By introducing a restriction of kitchen devices on Fridays, we implemented a conjunctive causal structure that is temporally limited to a particular weekday. This kind of causal structure allows us to control for the frequency of outcome and for cause effect co-variation across the two normality conditions described below.

2.3.2.2 Normality

In order to manipulate action normality more naturally, we used animated video clips. After having read the introductory text, participants proceeded to watch a clip that shows a week in the office, from Monday to Friday (ca. 40s). In the *'Both agents normal'* condition, both agents Agent 1 and Agent 2 use the kettles (microwaves) from Monday to Thursday every day. In the *'Agent 1 normal, Agent 2 abnormal'* condition, only Agent 1 uses a kettle [microwave] from Monday to Friday, and Agent 2 only on Friday. In both conditions, both agents use the devices on Friday *at the same time* and as a result, the power stops. In our design, Agent 1 acts as the fixed agent: Agent 1's action is always 'normal' or typical, while Agent 2 acts as varied agents, with the normality of Agent 2's action varying across conditions.

2.3.2.3 Causal Question

At the end of the clip, participants were asked to what extent they agree with the following two causal rating questions about the agents on 7-point Likert scales (1-'strongly disagree', 7-'strongly agree'): "Agent 1 [2] caused the power failure" (with one scale for each agent)[2], testing a graded notion of causality (Halpern &

[2] While the majority of studies in this area have focused on *intra-agent* comparisons (Icard et al., 2017; Kominsky et al., 2015), i.e. participants evaluate either fixed or varied agent only, our studies

Hitchcock, 2015). The order of the questions was randomised across participants.

After completing the clip and the causal agreement questions, participants had to answer one more comprehension check question concerning the frequency of the agents' action in the clip. After having watched both clips for each normality condition and responded to the causal rating questions, they proceeded to a final question about the agents' epistemic states.

2.3.2.4 Expectation Question

In the final part of the experiment, participant had to rate their agreement about the agents' epistemic states in both clips. This question served as a control question to assess whether the difference in the agents' action frequency also corresponded to a difference in people's judgments about the agents' expectations about each other's actions. We used retrospective ratings of the agents' expectations about the other's actions as a proxy for how likely the agent was to foresee the consequences of their action. People were asked to rate their agreement with two statements for each of the two video clips they saw on 7-point Likert scale (1-'strongly disagree', 7-'strongly agree'): "Agent 1 (2) expected Agent 2 (1) to use the kettle (microwave) on Friday.". The order of the questions was randomized.

Participants completed the experiment by providing demographic information. On average, it took participants 11 minutes (*SD* = 10.14) to complete Experiment 1.

2.3.3 Results

2.3.3.1 Causal Rating

We analysed participants' agreement with the causal statements by comparing a series of linear mixed models using the `lme` package `lme` package in `R`, with participants as random effects.

The analysis revealed a significant main effect for agent, $\chi^2(1) = 39.99$, $p < .001$, $R^2_c = .22$, normality $\chi^2(1) = 34.68$, $p < .001$, $R^2_c = .28$, and a significant interaction of normality and agent $\chi^2(1) = 49.67$, $p < .001$, $R^2_c = .35$.[3] Analysing

employ an *inter-agent* comparison design. We let participants rate both fixed [Agent 1] and varied [Agent 2] agent and compare the ratings of both agents in each experimental condition. See Sytsma (2019b) for a systematic review of effects of inter vs. intra-agent comparison contrasts.

[3] As effect size, we report the conditional R-squared for a full mixed linear effect model. R^2_c

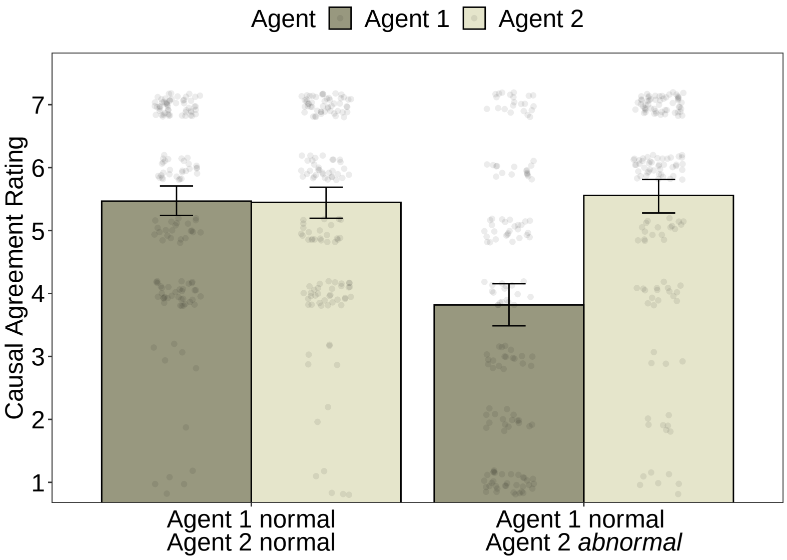


Figure 2.2: Experiment 1: Causal Ratings. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility.

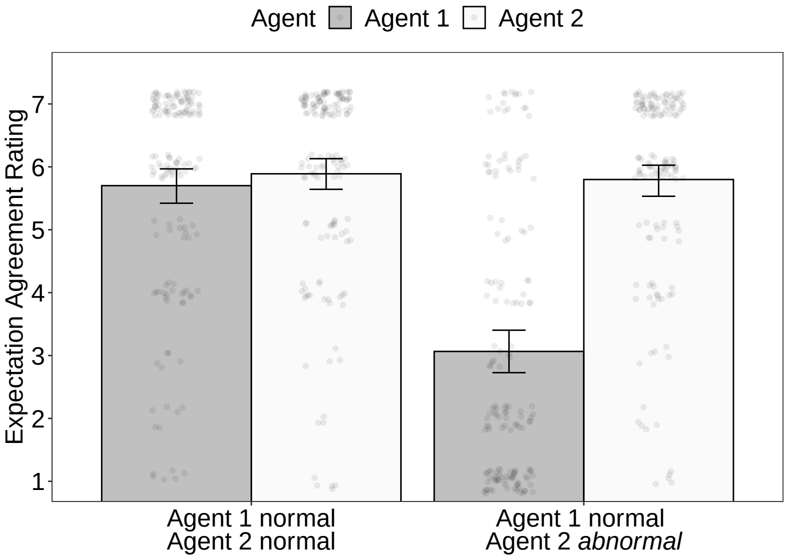


Figure 2.3: Experiment 1: Expectation Ratings. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility.

the effect of agent for both normality conditions showed that people judge the agent who has not acted frequently before (*M* = 5.56, *SD* = 1.68, 95% CI [5.29, 5.82]) as

provides the variance explained by a model including both fixed effects and random effects. R^2_c values for mixed-effects models are calculated using the r.squaredGLMM function of the MuMIn package (Barton & Barton, 2015) that implements a method developed by Nakagawa and Schielzeth (2013).

more causal than the frequently acting agent (*M* = 3.82, *SD* = 2.11, 95% CI [3.48, 4.15]), $t(465) = 10.10, p < .001$ (see Figure 3.2). When both agents act frequently, there is no difference between Agent 1 (*M* = 5.47, *SD* = 1.53, 95% CI [5.23, 5.71]) and Agent 2 (*M* = 5.45, *SD* = 1.54, 95% CI [5.21, 5.69]), $t(465) = .23, p = .91$. The interaction of normality and agent was independent of the scenario type, $\chi^2(4) = .63, p = .96$.

2.3.3.2 Expectation Rating

The analysis of expectation ratings revealed a significant main effect for agent $\chi^2(1) = 89, p < .001, R_c^2 = .25$, normality $\chi^2(1) = 94.55, p < .001, R_c^2 = .39$, and a significant interaction of normality and agent $\chi^2(1) = 101.85, p < .001, R_c^2 = .51$ (see Figure 3.3).

Analysing the effect of agent for both 'normality' conditions showed no significant difference between the varied (*M* = 5.89, *SD* = 1.51, 95% CI [5.64, 6.13]) and fixed agent (*M* = 5.70, *SD* = 1.69, 95% CI [5.43, 5.64]) when both act frequently $t(465) = 1.11, p < .01$, and crucially, a difference between the frequently (*M* = 3.06, *SD* = 2.16, 95% *CI* [2.72, 3.40]) and the infrequently acting agent (*M* = 5.80, *SD* = 1.58, 95% CI [5.55, 6.05]), $\chi^2(1) = 129.97, p < .001$. There was a significant interaction effect of the scenario type, $t(465) = 16.16, p < .001, R_c^2 = .02$. A decomposition of effects shows that the reduction in causal attribution to the frequently acting agent in the 'abnormal' condition is lower in the 'kettle' (*M* = 2.56, *SD* = 1.91, 95% CI [2.21, 3.10]) vs. 'microwave' scenario (*M* = 3.43, *SD* = 2.30, 95% CI [2.93, 3.93]), $\chi^2(4) = 6.21, p = .04$.

2.3.4 Discussion

Experiment 1 replicated the influence of action typicality on causal attributions to agents in a conjunctive causal structure. People judge a difference in the causality of two agents if these agents differ in how often they have performed the causal action. More precisely, the agent who has not performed this action before, i.e. who acts atypically or 'abnormally', is judged as more causal for the outcome. While previous literature has demonstrated an increase in causal attributions to the abnormal agent, we find in our experiment that people reduce their causal attributions to the normal agent in order to express a difference in perceived causality between abnormal and normal agent (see "causal supersession" Kominsky et al., 2014). By using animated video clips, we were able to show this effect when the 'normality' or frequency of actions is manipulated in a sequential manner. Crucially, we found that the manipulation of action normality also has an impact on how people judge the agents' epistemic states. In hindsight, people judge that the abnormally acting agent expected their co-worker to act on Friday to a greater extent than vice-versa.

Experiment 1 confirmed the influence of abnormality on causal attributions, but also showed that this difference in causal judgments corresponds to a perceived difference in the agents' expectations about each other. This raises the question of what is driving people's causal perception of abnormal actions. Does statistical abnormality influence causal judgments, with epistemic states merely being a by-product, or is the epistemic state pivotal for people's causal attributions? The difference in action frequency corresponds to an epistemic advantage for the agent who usually does not engage in the respective action. At the most basic level, this epistemic advantage consists of foreseeing that one's action will cause the outcome. Experimental paradigms manipulating statistical normality hence often co-manipulate the agents' epistemic states. In the second experiment, we therefore wanted to test whether normality influences causal judgment when the agents do not know about each other's actions.

2.4 Experiment 2: Abnormality without Knowledge about Each Other

In Experiment 2, we aimed to investigate the influence of normality of actions on people's causal judgments when the normality of actions does not change the agents' epistemic states.

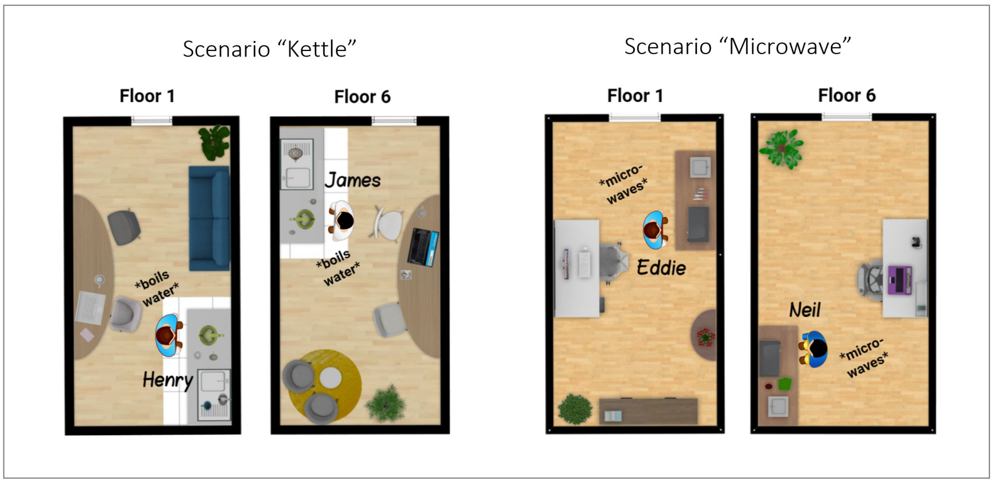


Figure 2.4: Illustration of two scenes from the video clips in Experiment 2. In each scenario, two agents work in separate offices on different floors. Depending on the scenario condition, each of the two offices has a kettle, or a microwave that the agents can use. In the scenario pictured, both agents use the item at the same time ('both agents normal' condition).

2.4.1 Participants and Design

For Experiment 2, 149 participants were recruited via Amazon Mechanical Turk. 19 participants were excluded for failing five or more out of ten comprehension check questions (see Appendix A.2), leaving a final sample size of $N = 130$ ($M_{\text{age}} = 36.15$, $SD_{\text{age}} = 10.81$, $N_{\text{female}} = 43$).[4] As in Experiment 1, we adopted a 2 normality ("Both agents normal" vs. "Agent 1 normal, Agent 2 abnormal") × 2 agent (Agent 1: fixed agent vs. Agent 2: varied agent) × 2 scenario ("kettle" vs. "microwave") mixed design. The factors 'normality' and 'agent' were manipulated within participants, 'scenario' was manipulated between participants.

2.4.2 Material and Procedure

The material and procedure closely followed Experiment 1, with one crucial difference regarding the agents' knowledge about each other. The two co-workers work in separate offices on different floors, and the participants were informed that the

[4] We performed a power analysis using the 'SimR' package (Green & MacLeod, 2016) based on the effect size estimates from Experiment 1. Our Experiment 2 with $N = 130$ had an observed power of 1 CI [96.3; 100] to detect a significant interaction of normality × agent for both causal and expectation judgments at $p < 0.05$.

agents do not know each other and have never met (Figure 2.4). As in Experiment 1, the company has introduced ‘Green Friday’ on which the use of more than one kettle [microwave] in the offices will lead to a power cutout in the company. Frequency of actions was manipulated as in Experiment 1, with both agents being located in separate offices. On Friday, both agents make use of the kettle [microwave] and a power failure occurs.

2.4.2.1 Cause and Expectation Question

At the end of each clip, participants were asked to what extent they agree with the following two questions about Friday on 7-point Likert scale (1-‘strongly disagree’, 7-‘strongly agree’): “Agent 1 (2) caused the power failure.” The order of the questions was randomised across participants. At the end of the experiment, participants had to rate their agreement about the agents’ epistemic states in both clips. We kept the epistemic rating question from Experiment 1 as a control question for our manipulation. People were asked to rate their agreement with two statements for each of the two video clip they saw on 7-point Likert scale (1-‘strongly disagree’, 7-‘strongly agree’): “Agent 1 (2) expected Agent 2 (1) to use the kettle (microwave) on Friday.” The order of the questions was randomised.

2.4.3 Results

2.4.3.1 Causal Rating

A Mixed Linear Model analysis on causal ratings revealed no significant main effect for agent, $\chi^2(1) = 0.21$, $p = .64$, normality $\chi^2(1) = 2.81$, $p = .09$, nor for the interaction between normality and agent $\chi^2(1)= .29, p = .59$. The infrequently agent (M = 4.88, SD = 1.85, 95% CI [4.57, 5.20]) is judged equally causal as the agent who has acted frequently before (M = 4.78, SD = 1.93, 95% CI [4.45, 5.11]) (see Figure 3.5). There was no significant interaction effect with scenario type, $\chi^2(4) = 7.65, p = .11$.

2.4.3.2 Expectation Rating

There was a significant main effect for agent $\chi^2(1) = 4.80$, p= .03, R_c^2 = .81, normality $\chi^2(1) = 11.90, p < .001$, R_c^2 = .81, and for the interaction between normality

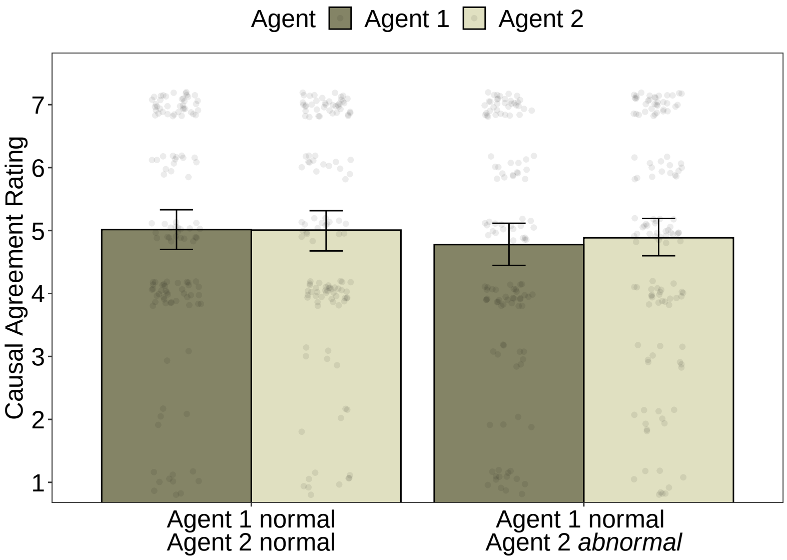


Figure 2.5: Experiment 2: Causal Ratings. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility.

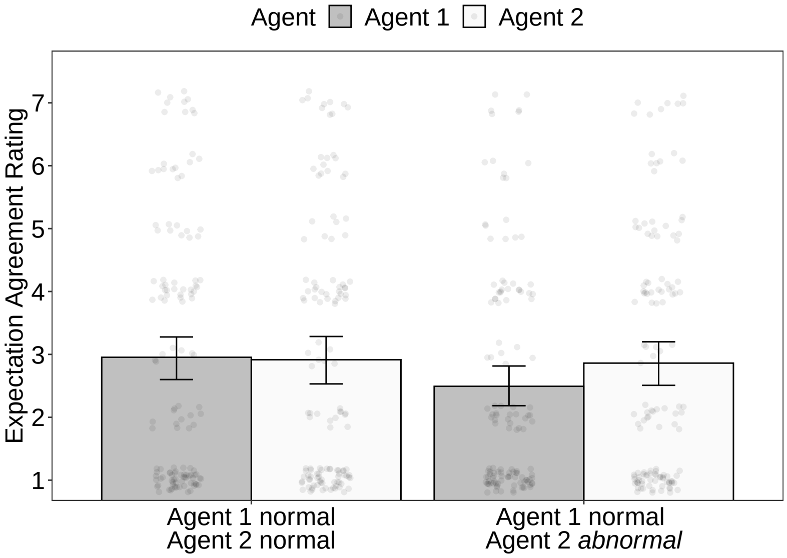


Figure 2.6: Experiment 2: Expectation Ratings. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility.

and agent $\chi^2(1) = 7.64$, $p < .001$, $R_c^2 = .82$.

While overall disagreeing that (M $<$ 4) the agents expected each other's behaviour, people disagree slightly less with the statement that the frequently acting agent expected the infrequently acting co-worker to act (M = 2.49, SD = 1.82, 95%

CI [2.18, 2.81]), than vice versa (*M* = 2.86, *SD* = 1.96, 95% CI [2.52, 3.20]), *t*(393) = -3.54, $p < .001$ (see Figure 3.6). We also found a significant three way interaction of normality, agent and scenario type $\chi^2(4) = 10.84$, $p = .02$, $R^2_c = .83$. Decomposing the three-way interaction revealed that in the 'both agents normal' condition, participants disagree less with statements about the agents' expectations in the kettle scenario (*M* = 3.27, *SD* = 2.11, 95% CI [2.92, 3.63]) compared with the microwave scenario (*M* = 2.56, *SD* = 1.96, 95% CI [2.22, 2.91]), $\chi^2(1) = 4.02$, $p = .04$.

2.4.4 Discussion

Experiment 2 showed that the normality of actions does not influence people's causal judgements when the agents do not know about each other. An abnormally acting agent is judged equally causal for the outcome as a normally acting agent when both are unaware of each other's actions. Experiment 2 suggests that people do not consider the abnormality of an agent's action for causal judgments if the abnormal agent does not have an advantage in foreseeing the outcome. In general, causal ratings in Experiment 2 were slightly lower than in Experiment, but still above mid-point ($M > 4$), indicating that people generally do judge the agents to be causal. In light of the fact that participants still identified the agents as causal for the outcome, we take this as evidence that the lack of a difference between the normal and abnormal agent is not due to a lack of causal attribution in general. Rather, participants do not perceive a causal difference between the two agents because there is no epistemic difference.

While participants overall tended to disagree with the statement that each agent expected their colleague to perform the causal action on Friday, there was a very small but significant difference between the agents in the abnormality condition. In general, retrospective evaluation of the agents' expectations allows people to attribute certain knowledge states to the agents in hindsight in order to make sense of their behaviour. One way to address this potential issue is to gather people's on-line judgements about the agents' knowledge about each other prior to the final outcome scenario. In Experiment 1 and 2, we investigated whether retrospective epistemic ratings correspond to the causal judgments that participants have given

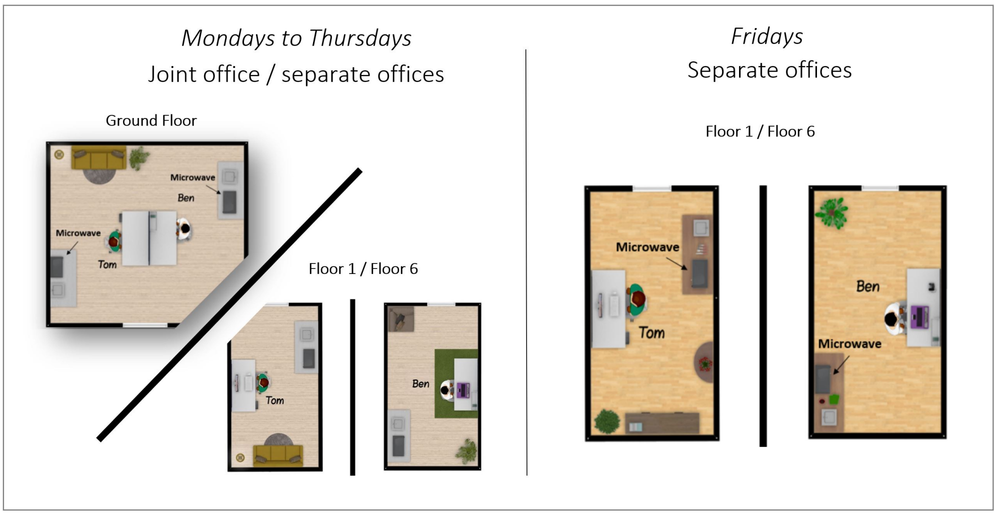


Figure 2.7: Illustration of the two experimental conditions of the 'microwave' scenario in Experiment 3. The two agents start off the week working either in a joint office ('knowledge condition') or in two separate offices on different floors ('no knowledge condition'). On Friday, in both knowledge conditions, the agents move into two different offices on separate floors.

before. In Experiment 3, we wanted reverse the order of these two rating types, investigating whether people's prospective evaluation of the agents' epistemic states prior to the outcome corresponds to their causal judgments after the outcome has occurred. In addition, we were interested in the scope of influence of epistemic states. For Experiment 3, we wanted to test whether the asymmetry in expectations also influences causal judgements when the agents' expectations about each other need to generalise to a novel context.

2.5 Experiment 3: Do Epistemic States influence Causal Judgments about Agents in Novel Contexts?

In Experiment 3, we tested whether an asymmetry in agents' epistemic states influences people's causal judgments even if the agents' expectation about each other's actions have to be transferred to a novel context.

2.5.1 Participants and Design

We recruited 145 participants on Amazon Mechanical Turk. Three participants were excluded for failing five or more out of ten comprehension check questions (see Appendix A.3), leaving a final sample size of $N = 142$ (M_{age} = 37.09, SD_{age} = 10.64, N_{female} = 58). In Experiment 3, we adopted 2 knowledge (knowledge about each other vs. no knowledge about each other) × 2 agent (Agent 1: normal vs. Agent 2: abnormal) × 2 scenario ("coffee machine" vs. "microwave") mixed design. We replaced the kettle with a coffee machine. The factors 'knowledge' and 'agent' were manipulated within participants, 'scenario' was manipulated between participants.

2.5.2 Material and Procedure

The material of Experiment 3 closely matched the paradigm of Experiment 1 and 2. Participants read an introduction to the scenario, and completed a series of comprehension check questions. However, the scenario in Experiment 3 included a change in the office set up to allow for the manipulation of the agents' epistemic states (see Figure 2.7).

2.5.2.1 Normality & Knowledge

In Experiment 3, the agents need to move into two separate office spaces on different floors on Friday because their usual office is needed for meetings. We varied whether their usual office space from Mondays to Thursdays was a joint office, or whether they worked in separate offices. In the *knowledge condition*, the agents work together in one office from Monday to Thursday, and therefore know whether the other agent is using the respective device each day. In the *no knowledge condition*, the agents work in different offices on separate floors, and do not know about each other. Agent 1 uses the device from Monday to Friday ("Normal Agent") and Agent 2 only uses the device on Friday ("Abnormal Agent").

2.5.2.2 Causal Structure

In line with the company policy, the use of two coffee machines [microwaves] on Fridays will lead to a power failure on Friday.

After having read the introduction and completed four comprehension check questions, participants proceeded to watch a first video clip. In this video clip, the weekdays from Monday to Thursday are shown, and on each of these day, only one of the two agents uses the coffee machine [microwave]. Depending on the knowledge condition, the agents are in a joint office space or in separate offices.

2.5.2.3 Expectation Rating

After having watched the first clip, participants are reminded that the agents need to move out of their current office and move into two separate offices the next day, Friday. They are then asked to rate the following statements about the agents' epistemic states on a 7-point Likert scale (1-'strongly disagree', 7-'strongly agree'): "Agent 1 (2) expects Agent 2 (1) to use the coffee machine (microwave) on Friday." As in the experiments before, the rating about what the agents expect about each other function as a proxy for how likely they think they might cause the outcome.

Participants then proceeded to watch a second video clip about the following Friday on which both agents have moved into their new offices. On this day, both of them use the coffee machine [microwave], and a power failure occurs.

2.5.2.4 Causal Rating

Participants were then asked to what extent they agree with the following two questions about Friday on 7-point Likert scale (1-'strongly disagree', 7-'strongly agree'): "Agent 1 (2) caused the power failure." After having completed the causal judgment task, participants had to answer one more manipulation check question about the action frequency of both agents (see Appendix A.3).

2.5.3 Results

2.5.3.1 Expectation Ratings

A Mixed Linear Model analysis on expectation ratings revealed a significant main effect for the factor agent (Agent 1: normal vs. Agent 2: abnormal), $\chi^2(1) = 196.1$, $p < .001$, $R_c^2 = .29$, knowledge $\chi^2(1) = 54.45$, $p < .001$, $R_c^2 = .36$, and a significant interaction for knowledge and agent $\chi^2(1) = 212.91$, $p < .001$, $R_c^2 = .60$.

When agents usually work in a joint office (Knowledge condition), participants judge the abnormal agent to expect the normal agent to act on Friday to a greater

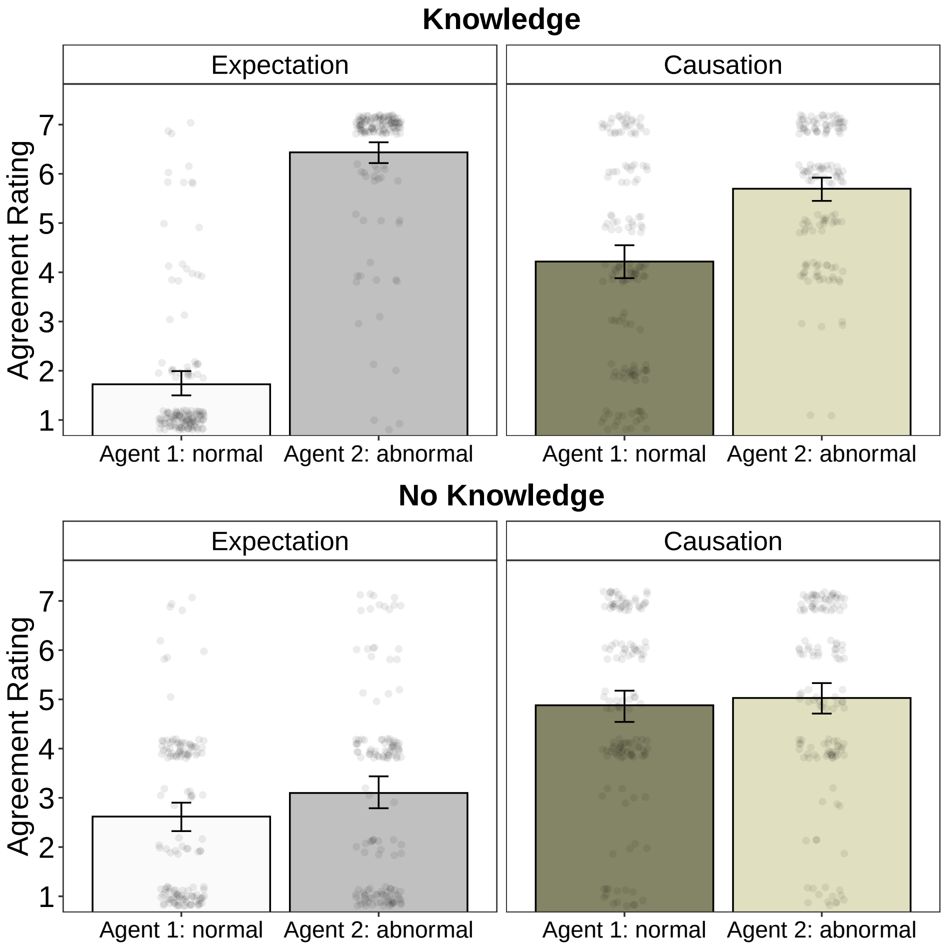


Figure 2.8: Experiment 3: Expectation and Causal Ratings. Agreement Ratings for the agents' epistemic states and causation of the outcome, depending on the 'knowledge' condition. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility.

extent (*M* = 6.44, *SD* = 1.30, 95% CI [6.22, 6.51]) than vice versa (*M* = 1.73, *SD* = 1.49, 95% CI [1.48, 1.97]), *t*(429) = 26.02, $p < .001$ (see Figure 2.8). When both agents usually work in separate offices, agreement ratings are reduced and people agree slightly more about the prospective expectations of the abnormal agent (*M* = 3.10, *SD* = 1.94, 95% *CI* [2.78, 3.42]) compared to the normal agent (*M* = 2.62, *SD* = 1.66, 95% CI [2.35, 2.89]), *t*(429) = 2.65, $p < .01$. There was no interaction with scenario type, *t*(429) = -27.03, *p* = .35.

2.5.3.2 Causal Ratings

The analysis of causal judgments revealed a significant main effect for the factor agent $\chi^2(1) = 44.84$, $p < .001$, $R_c^2 = .44$ and a significant interaction for knowledge and agent $\chi^2(1) = 32.98$, $p < .001$, $R_c^2 = .48$.

The abnormal agent is judged as more causal (M = 5.70, SD = 1.34, 95% CI [5.48, 5.92]) than the normally acting agent (M = 4.22, SD = 2.08, 95% CI [3.88, 4.56]) when both agents usually work in a joint office and know each other well, $t(429) = 9.16$, $p < .001$ (see Figure 2.8). In contrast, when the agents work in separate offices, people judge no causal difference between the abnormal (M = 5.03, SD = 1.87, 95% CI [4.72, 5.34]) and normal agent (M = 4.88, SD = 1.88, 95% CI [4.57, 5.19]) in causing the outcome on Friday, $t(429) = 0.92$, $p = .23$. There was no interaction with scenario type, $\chi^2(4) = 5.86$, $p = .21$, $R_c^2 < .01$.

2.5.4 Discussion

In Experiment 3, we gathered more evidence for our hypothesis that people's causal attributions to atypical actions are driven by the agents' epistemic states. The results show that in case of mutual knowledge about each others' habits, the abnormal agent is judged to expect a future action of the typically agent to a greater extent than vice versa. The abnormal agent is subsequently judged as more causal than the normal agent, even when the two causal actions occur in a context in which both agents need to predict the other's behaviour based on what they have learned about each other in the past. In contrast, if there is no such prior expectation, people do not perceive a causal difference between the abnormal and normal agent. Experiment 3 confirms our hypothesis in a paradigm in which the order of causal and epistemic ratings is reversed. Crucially, it shows that causal judgments can be influenced by agents' epistemic states in cases in which prior expectations generalise to novel contexts.

2.6 Mental State Inference from (Ab)normal Behaviour

In Experiment 1 - 3, we show that causal judgments are influenced by the normality of the agents' actions, but only if the difference in action normality is paralleled by an epistemic advantage for the abnormal agent. A typical or statistically normal action is more predictable than an abnormal action. If the consequence of one's action is co-dependent on that of another person, the relative foreseeability of the outcome for an agent is higher when the second acts frequently, rather than rarely or atypically. At minimum, this enables the abnormal agent to foresee the consequences of their action to a greater extent than the normal agent.

While Experiments 1 - 3 demonstrated the agents' asymmetry in expectations about each other, we have yet to show that this asymmetry also translates into the predicted difference in foreseeability of the outcome. This was the first aim of Experiment 4. However, given that the abnormal agent acts despite being able to foresee a negative outcome, this raises further questions about the agent's mental state. More precisely, acting atypically evokes inferences about mental states that go beyond the agent's will or desire to use the microwave or coffee machine. We hypothesize that in such a scenario, people will make inferences from the (a)typicality of agents' overt behaviour about their "outcome"-oriented mental state, e.g. their intention or desire towards the outcome (C. Baker, Saxe, & Tenenbaum, 2011; C. L. Baker, 2012; Saxe & Houlihan, 2017). Given the asymmetry in epistemic states, we predict that an agent deviating from their usual routine of (non)-action will lead people to infer an increase in the degree of this agent's outcome-directed mental state, like a desire or intention to cause a power failure (Jones & Davis, 1965; Jones et al., 1961).

2.6.1 Bayesian Network Model of Mental State Inference

Our hypothesis can be formalised using a causal Bayesian network Model (see Figure 2.9). A Bayesian network (Pearl, 2009) is a formalism that uses a directed acyclic graph to represent the probabilistic dependencies between variables. The

Table 2.1: Conditional Probability Tables for ‘Knowledge’ condition: Conditional probability tables for the variables ‘Normality’, ‘Mental State’, ‘Expectation about Agent 2’ and ‘Action’ of agent A1. The probability tables are symmetrical for agents A1 and A2.

(a) ‘Normality’

Normality A1	
true	0.8
false	0.2

(b) ‘Mental State’

Mental State A1	
true	0.1
false	0.9

(c) ‘Expectation about A2’

A1’s Expectation about A2		
Normality of A2	*false*	*true*
false	0.9	0.1
true	0.1	0.9

(d) ‘Action’

Action of A1								
Mental State	*true*				*false*			
Normality of A1’s Action	*true*		*false*		*true*		*false*	
Expectation about A2’s Action	*true*	*false*	*true*	*false*	*true*	*false*	*true*	*false*
true	0.9	0.8	0.9	0.2	0.1	0.8	0.1	0.2
false	0.1	0.2	0.1	0.8	0.9	0.2	0.9	0.8

Table 2.2: Conditional Probability Tables for ‘No Knowledge’ condition: Conditional probability table for the variables ‘Expectation about Agent 2’ of agent $A1$. All other variables keep their values from the ‘Knowledge’ condition (see Table 1). The probability tables are symmetrical for agents $A1$ and $A2$.

(a) ‘A1’s Expectation about A2’

A1’s Expectation about A2	
false	0.9
true	0.1

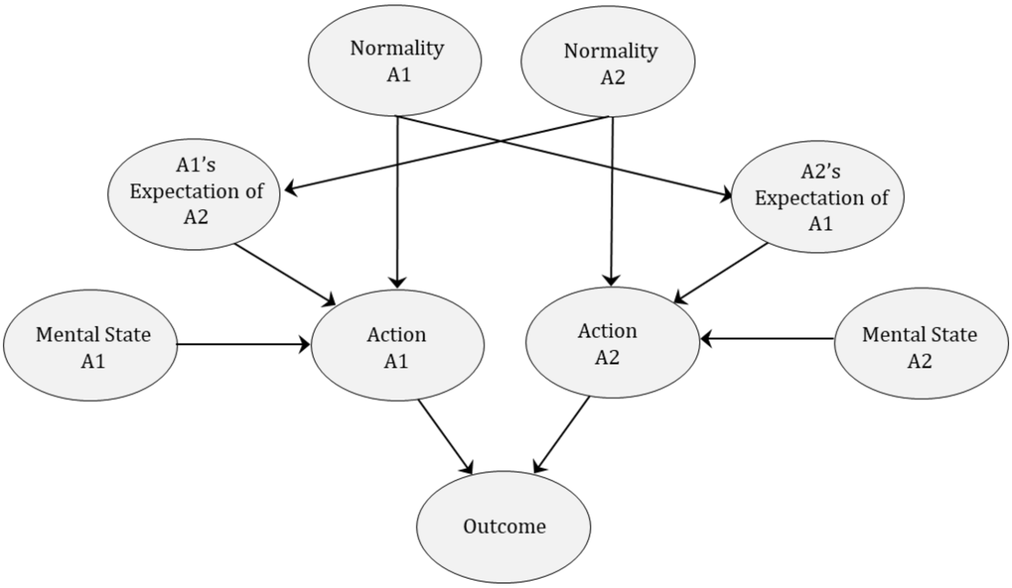


Figure 2.9: Model 'Knowledge'. Bayesian network Model for a conjunctive causal structure in which the outcome depends on the actions of two agents "A1" and "A2". Whether an agents acts depends on the normality of their action, their expectation about the other agent, and their mental state towards the outcome. In the "knowledge" condition, there is a link between the normality of an an agent's action is, and whether the other co-agent expects this agent's action.

qualitative side of a Bayesian network is the graph structure, where a link from X to Y corresponds to the claim that Y depends on X. For the quantitative side, each variable has a conditional probability table (CPT) that specifies the probability of that variable given the possible values of its immediate causes (its parents in the graph). Variables with no parents are assigned prior probabilities.

When we acquire evidence about any of the variables in the model, we can use Bayes' rule to update the probabilities of the other variables. Bayesian networks are hence ideal for diagnostic inference: given some observed effect we can infer the probabilities of the possible causes of this effect. Here, we use Bayesian networks to model our scenarios, aiming to capture the key causes of the agents' actions, and how their probabilities should be updated given the manipulations in our experiments. Assuming that an agent's action is influenced by the normality of their actions, what they know about others and, crucially, their mental state, the model allows us to predict people's degree of inference about an agent's goal-directed mental state based on the other two factors. We develop this model for a scenario

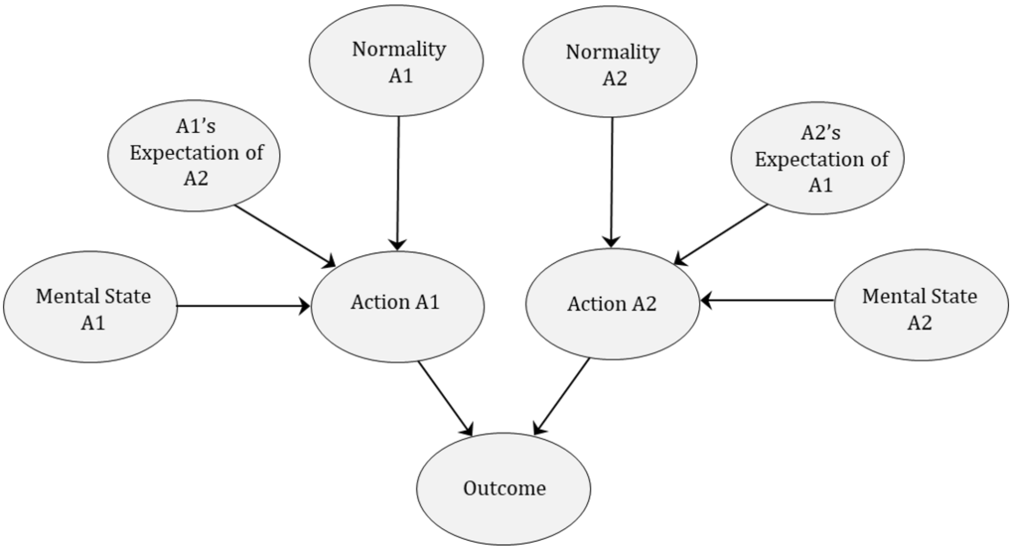


Figure 2.10: Model 'No Knowledge'. Bayesian network Model for a conjunctive causal structure in which the outcome depends on the actions of two agents "A1" and "A2". In the "No Knowledge" condition, there is no link between how normal an agent's action is, and whether the co-agent expects this agent's action

in which the two agents know about each other (see Figure 2.9), and a scenario in which neither agent knows about the other (see Figure 2.10).

2.6.1.1 Graph Structure

Central to our model is the assumption that an agent's action is influenced by three causal variables: how typical or "normal" their action is ("Normality"), what the agent expects other agents to do ("Expectation"), and finally, what kind of mental state the agent is in, that is, whether the agent intends or desires the outcome ("Outcome-Oriented Mental State"). These are represented as causal parents of the agent's action, and the same template is used for each agent, A1 and A2. In the "*Knowledge*" version of our model, whether A1 expects A2 to act is influenced by the normality of A2's behaviour (see Figure 2.9). In order to capture the assumption that normality of behaviour influences expectations, we add links from the normality of one agent's behaviour to the other agent's expectation about that behaviour. In the "*No Knowledge*" version of our model, the variable "A1's Expectation of A2" is independent of the normality of A2's behaviour (see Figure 2.10). Finally, whether the outcome in a conjunctive structure occurs depends on the action of both agents

A1 and A2.

2.6.1.2 Parametrising the Model

We parameterise the network with illustrative probabilities. The conditional probabilities for each variable in this model serve as a rough approximation for our hypotheses, and can be flexibly adapted. For a start, we determine the probability for abnormal behaviour based on an agent acting on one out of five days, $P(A1\ abnormal) = 1/5 = 0.2$ (see Table 2.1a). Furthermore, in line with expectation ratings from Experiment 3, we assume that in the "Knowledge" condition, $A1$'s frequent, "normal" action will lead $A2$ to expect A1 to act with a likelihood of 90%, $P(A1`s\ Expectation\ about\ A2 \mid A2\ normal) = 0.9$ (see Table 2.1c). In the condition in which the agents do not know about each other ("*No Knowledge*"), we have set the prior probability of an agent knowing about the other to 10%, $P(A1\ Expectation\ about\ A2) = 0.1$. (see Table 2.2). In our study, the outcome in the conjunctive structure is a power failure. For an outcome with negative valence, we assume that the prior probability of having a certain mental state, e.g. a desire for bringing about a negative outcome, is low, $P(Mental\ State) = 0.1$ (see Table 2.1b) (see for example Chee & Murachver, 2012; Maselli & Altrocchi, 1969).

How does an agent's mental state with regards to a particular outcome influence their action? While there are various ways to model this influence, we make two basic stipulations. First, the probability of an agent acting who intends or desires the outcome and expects the other agent to act is high, i.e. $P(Action\ A1 \mid Mental\ State\ A1,\ A1\ expects\ A2) = 0.9$. (see Table 2.1d). Crucially, this is independent of how normal or abnormal the action is. Second, if the the agent intends the outcome but does not expect the other agent to act, we assume that the probability of this agent acting is simply how "normal" their action is, $P(Action\, A1 \mid Mental\ State\, A1, \neg\, A1\ expects\, A2) = P(Normality\ of\ A1)$.

2.6.2 Predictions

We now can compute the probability that an agent has an outcome-directed mental state given that both agents have acted in the 'knowledge' vs. 'no knowledge' con-

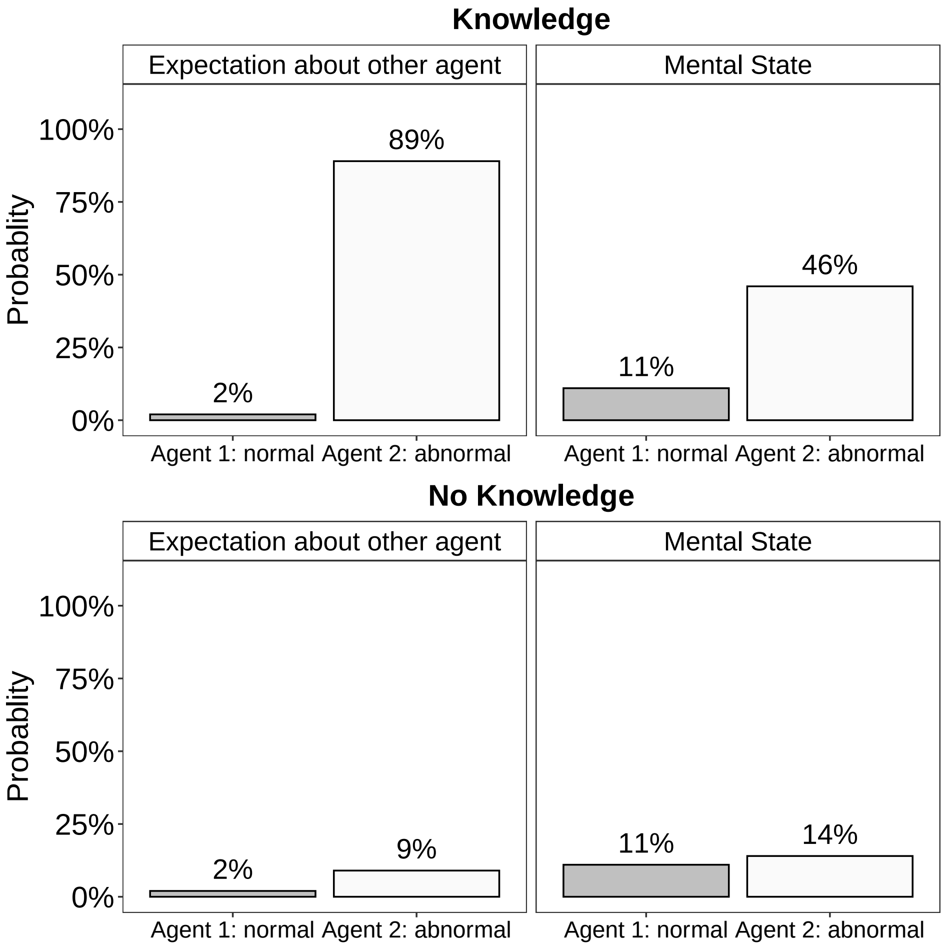


Figure 2.11: Bayesian Model predictions in the "Knowledge" and "No Knowledge" scenario. Bayesian model predictions for the probability of the normal [abnormal] agent i) expecting the abnormal [normal] to act, and ii) to have an outcome-directed mental state.

dition using Bayesian updating. For our purposes, we will use a model that has set the variable "normality" to 'true' for $A1$, the normal agent, and 'false' for $A2$, the abnormal agent.

2.6.2.1 Knowledge

When both agents act but differ in how normal their actions are, our model predicts the abnormal agent to expect the normal agent's action with a probability of $P(A2\ expects\ A1 \mid A1\ normal,\ A2\ abnormal,\ Action\ A1{,}Action\ A2) = 0.89$, and the probablity of the normal agent expecting the abnormal agent's action as $P(A1\,expects\,A2\,|\,A1\,normal{,}A2\,abnormal,\,Action\,A1{,}Action\,A2) = 0.02$ (see Fig-

ure 2.11 “Knowledge”). The probability of the normal agent having an outcome-directed mental state in such a scenario is
$P(Mental\ State\ A1 \mid A1\ normal,\ A2\ abnormal,\ Action\ A1, Action\ A2) = 0.11.$
In contrast, the probability for the abnormal agent to have an outcome-directed mental state is $P(Mental\ State\ A2 \mid A1\ normal, A2\ abnormal,\ Action\ A1, Action\ A2) =$ 0.46.

2.6.2.2 No Knowledge

When neither agent knows about the other, our model predicts that the probability of the normal agent expecting the abnormal agent to act is 2%, and vice versa 9% (see Figure 2.11 “No Knowledge”). The probability for the abnormal agent having an outcome-directed mental state is 14%, and 11% for the normal agent .

2.6.2.3 Excursion: Positive and Neutral Outcomes

In contrast to the agent’s expectations, the degree of predicted inference about an outcome-oriented mental states is to some extent dependent on their prior probability. If we would consider a case in which the outcome is positive, we would assume that the probability for having a mental state oriented at a positive outcome is high, e.g. $P(Mental\ State) = 0.9$. In such a case, our model predicts that the inferred probability of agents possessing this mental state is much higher, and the inferred difference smaller, 91% for the normal agent and 99% for the abnormal agent.

There is a lot of flexibility in how to specify the probabilities of each variable in this model, for example the assumed prior probability of having a mental state oriented at the outcome or how likely an agent with an outcome-oriented mental state is to act despite them usually not performing the causal action. Hence, exact values are not as important as comparative values. Crucially, the model in our example renders comparative differences between the mental states of the normal and abnormal agent. Based on this model, we predict that people are more likely to infer an outcome-directed mental state from abnormal behaviour compared to normal behaviour in case of knowledge. When neither agent knows about the other, inference about an outcome-directed mental state for both abnormal and normal agent will be equally low.

2.7 Experiment 4: Inferences about Outcome-Oriented Mental States

In Experiment 4, we wanted to investigate people's inferences about agents' epistemic and outcome-oriented mental states based on the statistical normality of their actions and their knowledge about each other. First, we wanted to assess an intermediate step that has been tacitly assumed in our model, but not yet tested: that an asymmetry in knowledge about the other agent's action also corresponds to an asymmetry in expecting the outcome to occur as a result of one's action. At minimum, we expect people to attribute a higher expectation of the outcome to an abnormal compared to a normal agent when the agents know about each other, but not when they do not know about each other. Second, we wanted to put the qualitative predictions of the Bayesian network model about people's outcome-oriented mental state inferences to test.

2.7.1 Participants and Design

We recruited 163 participants on Amazon Mechanical Turk. One participant was excluded for failing five or more out of ten comprehension check questions (see Appendix A.4), leaving a final sample size of $N = 162$ (M_{age} = 37.62, SD_{age} = 11.41, N_{female} = 71)[5]. In Experiment 4, we adopted a 2 knowledge (knowledge about each other vs. no knowledge about each other) × 2 agent (Agent 1: normal vs. Agent 2: abnormal) × 2 scenario ("coffee machine" vs. "microwave") design. The factors 'knowledge' and 'agent' were manipulated within participants, 'scenario' was manipulated between participants.

2.7.2 Material and Procedure

In Experiment 4, we used the 'Agent 1: normal, Agent 2: abnormal' condition from Experiment 1 ("Agents know about each other") and Experiment 2 ("Agents don't know about each other"). As before, participants were introduced to the scenario, completed four manipulation check questions, and then proceeded to watch

[5] Power analysis based on the effect size from Experiment 3 showed that Experiment 4 with N = 162 had a power of 1 CI [96.3; 100] to detect a significant interaction of agent × knowledge on expectation judgments at $p < 0.05$.

an animated video clip. In this clip, one agent frequently uses a coffee machine [microwave] from Monday to Thursday (“Agent 1: normal”) while the other agent does not (“Agent 2: abnormal”). Both then cause a power failure by using both devices on Friday. Depending on the ‘knowledge’ condition, the agent work in one joint or two separate offices.

2.7.2.1 Epistemic Questions

After watching the whole video clip including the final day, participants were first asked to rate the agents’ expectations about each other: “Agent 1 (2) expected Agent 2 (1) to use the coffee machine (microwave) on Friday.” (7 - point Likert scale; 1 - ’strongly disagree’, 7 - ’strongly agree’). The order of the two questions was randomised across participants. Participants then completed a second follow up question: “Given your answer to [question] (1), Agent 1 (2) _______ the outcome. ” The question was followed by a list of five outcome-oriented mental states, together with 7-point Likert scale (1 - ‘strongly disagree’, 7 - ‘strongly agree’) for each mental state: 1) “expected”, 2) “did not mind”, 3) “liked”, 4) “desired”, 5) “intended”. Liking, desiring and intending qualify as being dispositional for an outcome-directed action, i.e. provide the mental condition for acting towards a goal or outcome (Brandstätter, Lengfelder, & Gollwitzer, 2001; Kuhlmeier, Wynn, & Bloom, 2003; Perugini & Bagozzi, 2001; Ryle, 2009). We included the mental state of ‘being indifferent’ because, while not necessarily being dispositional for an action, being indifferent towards the outcome does not prevent an agent from acting despite foreseeing it.

Participants had to rate their agreement with the insertion of each outcome-oriented mental state into the statement. This question was asked for both agents, and the order of each rating set for the agents was randomised across participants. After having answered all mental state rating questions, participant proceeded to the comprehension check question about the agents’ action frequency in the video clip. Participants completed both the ‘knowledge’ and ‘no knowledge’ condition in randomized order.

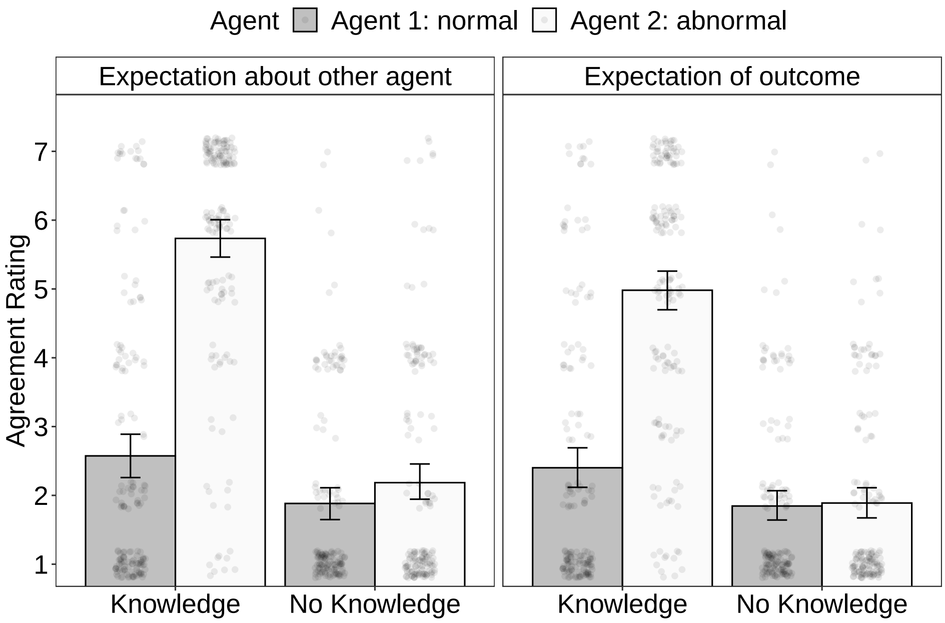


Figure 2.12: Experiment 4: Epistemic State Ratings. Ratings for the agent's expectation about i) the other agent and ii) the outcome are given with regards to the abnormal and normal agent. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility

2.7.3 Results

2.7.3.1 Expectation about Each Other

The analysis of agreement ratings with the statement "Agent 1 (2) expected Agent 2 (1) to use the coffee machine (microwave) on Friday." revealed a significant main effect for the factor agent $\chi^2(1) = 97.67, p < .001, R_c^2 = .14$, knowledge, $\chi^2(1) = 183.56, p < .001, R_c^2 = .38$, and a significant interaction for knowledge and agent $\chi^2(1) = 122, p < .001, R_c^2 = .51$.

The abnormal agent is judged to expect the other agent to act to a greater extent (*M* = 5.73, *SD* = 1.74, 95% CI [5.46, 6.00], $t(489) = 17.49, p < .001$, than the normally acting agent (*M* = 2.57, *SD* = 2.02, 95% CI [2.26, 2.89]) when both agents share an office, compared to when abnormal (*M* = 2.19, *SD* = 1.69. 95% *CI* [1.92, 2.45]) and normal agent do not know each other (*M* = 1.88, *SD* = 1.41, 95% CI [1.67, 2.10]), $\chi^2(1) = 9.70, p < .01$, (see Figure 3.15). There was no interaction with scenario type $\chi^2(4) = 2.11, p = .71$.

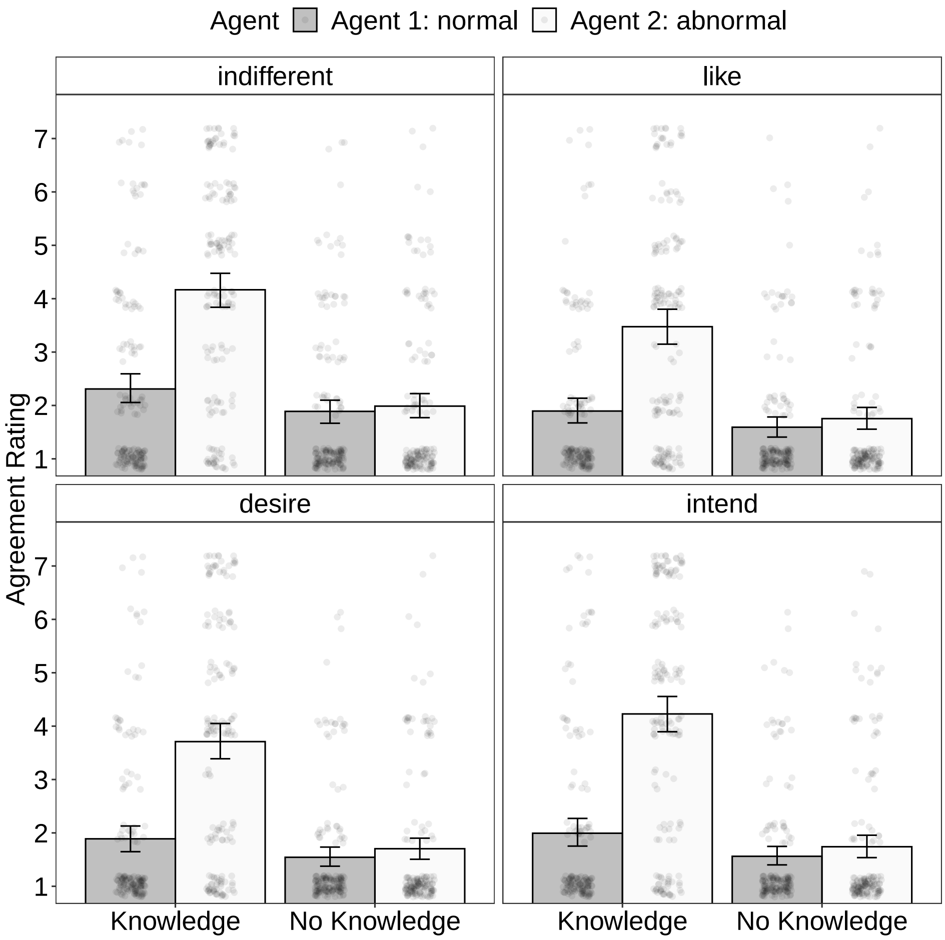


Figure 2.13: Experiment 4: Outcome-Oriented Mental State Ratings. Agreement ratings for four different mental states classes. Ratings for each outcome-oriented mental state are given with regards to the abnormal and normal agent. Error bars depict 95% Confidence Intervals. Grey dots are individual participants' judgments jittered for visibility.

2.7.3.2 Expectation about the Outcome

The analysis of the agreement ratings for expectation of the outcome revealed a significant main affect for agent, $\chi^2(1) = 65.92$, $p < .001$, $R^2_c = .09$, knowledge, $\chi^2(1) = 155.82$, $p < .001$, $R^2_c = .34$, and a significant interaction between agent and knowledge $\chi^2(1) = 98.08$, $p < .001$, $R^2_c = .46$.

When both agents know about each other, the abnormal agent was judged to expect the outcome to a greater extent ($M = 4.98$, $SD = 1.90$, 95% CI [4.69, 5.27]) than the normal agent ($M = 2.40$, $SD = 1.95$, 95% CI [2.10, 2.70]), $t(489) = 14.95$, $p < .001$. There is no difference when the agents do not know about each other, $t(489)$

= -.25, p = .80, (Abnormal agent: M = 1.89, SD = 1.40, 95% CI [1.67, 2.10]; Normal agent: M = 1.85, SD = 1.34, 95% CI [1.64, 2.05]) (see Figure 3.15). There was a significant interaction with scenario type $\chi^2(4)$ = 11.80, p = .02, R_c^2 = .47. Ratings for the abnormal agent in the 'no knowledge' were higher in the coffee scenario (M = 2.26, SD = 1.69, 95% CI [0.19, 1.69]) than in the microwave scenario, (M = 1.55, SD = 0.97, 95% CI [0.11, 0.96]), $t(119)$ = 3.22, $p < .01$.

2.7.3.3 Outcome-Oriented Mental State Attribution

We aggregated the four mental state classes 'being indifferent', 'liking', 'desiring' and 'intending' into one measure of outcome-oriented mental state attribution. The analysis of the agreement ratings for a mental state towards the outcome revealed a significant main affect for agent, $\chi^2(1)$ = 66.63, $p < .001$, R_c^2 = .25, knowledge, $\chi^2(1)$ = 119.97, $p < .001$, R_c^2 = .42, and a significant interaction between agent and knowledge $\chi^2(1)$ = 71.01, $p < .001$, R_c^2 = .50.

When the agents know about each other, the abnormal agent was judged to posses a greater degree of the outcome-oriented mental state (M = 3.90, SD = 1.97, 95% CI [3.59, 4.20]), than the normal agent (M = 2.02, SD = 1.51, 95% CI [1.79, 2.25]), $t(489)$ = -13.40, $p < .001$. People judge no difference when the agents do not know about each other, $t(489)$ = -1.07, p = .28, with the abnormal agent judged to have this mental state to the same extent (M = 1.80, SD = 1.33, 95% CI [1.59, 2.00]) as the normal agent (M = 1.64, SD = 1.14, 95% CI [1.47, 1.82]) (see Figure 3.16). There was no interaction with scenario type, $\chi^2(4)$ = 6.89, p = .14.

2.7.4 Discussion

Experiment 4 confirmed the qualitative predictions about outcome-oriented mental state inference from our Bayesian network model. First, in line with our previous studies, people judged the abnormal agent to expect the normal agent to act to a greater extent than vice versa, but only in the condition in which both agents know about each other. In addition, people also judged the abnormal agent to expect the outcome to a greater extent than the normal agent in the 'knowledge' condition, but not in the 'no knowledge' condition. Overall, participants inferred four types of outcome-oriented mental states – indifference, liking, desire and intention – to a greater extent from abnormal behaviour than from normal behaviour. However, this was only the case when both agents knew about each other.

The fact that epistemic state asymmetry also generates an asymmetry in inferences about outcome-oriented mental states raises the question whether causal attributions are partly driven by what people assume about the agents' intentions and desires, rather than just by the difference in foreseeability. In our studies, we have shown that "normality" or action typicality influences causal judgments by creating an epistemic asymmetry between two agents who act with different frequency. The argument we make here is compatible with this influence being mediated by further inferences about mental states. However, as we have speculated above, inferences about mental states can vary quite substantially depending on the valence of the outcome, while expectations of the outcome likely remain unaffected by this factor. Future studies will need to test the exact parameters in this model and their influence on causal attributions more rigorously.

2.8 General Discussion

The phenomenon that people systematically choose atypical and abnormal actions, agents, and objects as causes has been the subject of debate in philosophy as well as psychology (cf. Hart & Honoré, 1959/1985). Our preference for abnormal causes manifests in causal explanations and language (Gerstenberg & Icard, 2020; Hilton & Slugoski, 1986), causal judgments (Icard & Knobe, 2016; Knobe & Fraser, 2008; Kominsky et al., 2015), causal intervention (Cheng & Novick, 1991) and prospective causal thinking (Henne, O'Neill, Bello, Khemlani, & De Brigard, 2021). Several theories have been proposed to explain why we have a tendency to prefer atypical and abnormal factors as causes (Alicke et al., 2012; Cheng & Novick, 1991; Hilton & Jaspars, 1987; Hitchcock & Knobe, 2009; Icard et al., 2017; Samland & Waldmann, 2016; Woodward, 2001).

In this chapter, we did not aim to settle this debate. Rather, we have argued that, in terms of causal agents, an important and often overlooked factor driving people's causal attributions is the agents' epistemic states. In four experiments, we have shown that the tendency to judge an abnormally acting agent as more causal than a normally acting agent is influenced by the epistemic asymmetry between agents. People do not perceive a causal difference between an abnormal and normal agent if no epistemic asymmetry arises. In addition, we find that people are infer to a greater degree of an outcome-oriented mental state from an abnormal action, but again, only if this abnormal behaviour is accompanied by an epistemic advantage for the abnormal agent.

What are the implications of our findings? First, we discuss our results in the context of the current debate about norms in causal cognition. We will examine our findings in the light of prominent accounts explaining the influence of norms in causal cognition (Alicke & Rose, 2012a; Alicke et al., 2012; Henne, Niemi, et al., 2019; Icard et al., 2017; Knobe, 2009; Kominsky & Phillips, 2019; Samland et al., 2016a). Second, we will discuss how our findings relate to other research on the influence of abnormality on causal judgments, and by doing so, outline the limitations of our study and point out future directions of research. In particular, we will discuss research on atypically acting single agents (Kahneman & Miller, 1986), abnormal objects (Kominsky & Phillips, 2019) and a relatively novel finding in causal cognition research, the preference for normal causes in disjunctive structures (Gerstenberg & Icard, 2020; Icard et al., 2017).

2.8.1 Counterfactuals, Blame and Pragmatics

The influence of norms, or normality, on causal judgments has been the subject of recent debate in psychology and philosophy. Studies in this area show that agents who violate prescriptive or statistical norms receive higher causal attributions than norm-abiding agents (Icard & Knobe, 2016; Kahneman & Miller, 1986; Knobe, 2009; Kominsky & Phillips, 2019; Kominsky et al., 2015). Several accounts have been suggested to explain this pattern.

2.8.1.1 Counterfactuals

The *counterfactual reasoning* account (Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Knobe, 2009; Kominsky & Phillips, 2019; Kominsky et al., 2015) aims to explain the influence of norms and normality by reference to people's thinking about alternative possibilities. Normality is assumed to make certain counterfactual possibilities more relevant than others. A norm-violating causal action is mentally replaced by a norm-conforming action, hence highlighting the counterfactual dependence of the outcome on that particular action (Hitchcock & Knobe, 2009; Knobe, 2009; Kominsky et al., 2015; Phillips et al., 2015). Recent developments of this account have suggested an expectation-based account of normality (Kominsky & Phillips, 2019). According to this account, an action is only perceived as abnormal if the acting agent can expect their behaviour to be norm violating. This extension aims to explain why an agent who unknowingly violates a rule is not judged more causal than a norm-abiding agent (Samland & Waldmann, 2015, 2016).

How do our results perform in light of an expectation-based normality account? In terms of *group-level statistical normality* (Sytsma et al., 2012), an agent who acts in ignorance of what others do might not be aware which actions are frequently or typically performed, and is hence unaware of the atypicality of their own behaviour. In this sense, a counterfactual account would predict no causal difference between an abnormal and normal agent in the 'no knowledge' condition of our experiments. In terms of *agent-level statistical normality*, an agent can generally be expected to know when they deviate from their previous behaviour, independent of their knowledge about what others do. Whether an expectation-based account can explain the results in this chapter depends largely on how knowledge about statistical normality is conceptualised.

Following up on the idea by Kominsky and Phillips (2019), we think there is potential for counterfactual accounts to incorporate epistemic states more generally (Kirfel & Lagnado, 2017, 2018, 2019). Adopting a *theory of mind* account of counterfactual reasoning, people's simulation of counterfactual alternatives to an agent's action might depend on the agent's knowledge and beliefs broadly construed. An agent's ignorance about relevant features of their action might influence to what extent people consider alternative situations in which the agent does not perform this action. If the causal agent does not know, or could not have reasonably been assumed to know that they will cause harm or violate a rule by a certain action, people might not consider (or be less likely to sample) counterfactuals in which this action is undone. In this respect, however, people's counterfactual reasoning (Icard et al., 2017) or simulation of possibilities (P. N. Johnson-Laird, Khemlani, & Goodwin, 2015; S. S. Khemlani, Byrne, & Johnson-Laird, 2018) is *not* directly influenced by the statistical normality of actions. Instead, normality of action influences causal judgements by changing agents' epistemic states about (the consequences of) their action. Future research will need to test the influence of agents' epistemic states on counterfactual reasoning more directly.

2.8.1.2 Blame and Responsibility

A different class of theories aims to explain the role of prescriptive normality in causal cognition in terms of moral judgements. According to the *Culpable Control Model*, people's causal judgments are biased towards a desire to assign blame to the abnormal factor (Alicke, 2000; Alicke & Rose, 2012a). Increasing the perceived causal contribution to a norm-violating causal agent allows people to validate a spontaneous blame response. Sytsma et al. (2012) argue that people's ordinary concept of causation is itself normatively enriched, and that it is used similarly to the concept of responsibility (Sytsma, 2019a; Sytsma & Livengood, 2019; Sytsma et al., 2012). Samland et al. (2016a) suggest that norms shift people's pragmatic understanding of the verbal cause concept into the moral domain. In the context of norms and norm-violating agents, participants interpret the causal test question as a request to assign accountability (Samland & Waldmann, 2014, 2015, 2016).

The fact that action typicality influences causal judgments via epistemic states would allow blame-oriented accounts of causal judgments to extend their predictions to descriptive norms like action typicality. An agent is seen as more blameworthy, more responsible, or is held more accountable for a negative outcome if they could have foreseen the outcome (more) qua the typicality of their and other agents' actions. Sytsma (2019a) finds that an agent's character, such a being negligent, matters for causal attributions when this character trait warrants an inference about the agent's epistemic state about the outcome. In our studies, we show that people attribute high expectation of the outcome to the abnormal agent, yet are hesitant to attribute to them particularly strong intentions or desires for a negative outcome. In addition, according to our model, the degree of inferences about mental states like intentions and desires is sensitive to the nature of outcome, and might vary according to the perceived prior probability of an agent intending a good, bad, or neutral outcome. Studies, however, show a somewhat consistent difference in causal attributions to normal and abnormal causes for outcomes of different valence (Icard et al., 2017; Kominsky et al., 2015). In general, some blame-oriented accounts do not provide a fully fledged explanation of why people would attribute causality to an abnormal agent who causes a neutral or positive outcome (Icard et al., 2017). At the current stage, blame-oriented accounts are need to specify to what extent they take each of the epistemic and mental state components to influence attributions of blame, accountability or responsibility.

The findings from our experiments are in principle compatible with both accounts, and the studies in this chapter do not speak in clear favour for either of the two theoretical lines. The role of epistemic states for causal attributions provides interesting new challenges for both lines of research – accounts that assume counterfactual reasoning, as well as those that stipulate an underlying 'normative' judgment. An agent's epistemic state might influence reasoning about alternatives to their action, but they are also a crucial factor in blame and moral judgments. Further studies on how epistemic states determine causal attributions with the addition of response measures assessing blame and perceived norm-violations might help decide between these two accounts.

2.8.2 Single actions

The experiments in this chapter have exclusively focused on causal attributions to two agents in a conjunctive causal structure. As demonstrated by the 'hitchhiker case' (Kahneman & Miller, 1986), people are also more prone to generate counterfactual alternatives and attribute causality if a negative outcome results from a single agent performing an atypical vs. typical action (Fillon, Kutscher, & Feldman, 2020; Fillon et al., 2019; Hilton & Slugoski, 1986; Kahneman & Miller, 1986; Monroe & Ysidron, 2021). Crucially, in these scenarios, the action is not directly causal for the outcome, as the agent passively experiences an external event (e.g. car crash, robbery). This breaks the usual link between typicality of actions and foreseeability of outcome, and makes assumptions and inferences about the agent's epistemic state speculative. Kutscher and Feldman (2019) find that people's anticipation of the likelihood of a negative incident is the same for atypical compared to typical actions (Macrae, 1992; Turley, Sanna, & Reiter, 1995).

However, Macrae (1992) presumes that an agent who deviates from a routine behaviour for which negative incidents have been absent in the past risks greater uncertainty towards the consequences of their abnormal action than someone who abides with past behaviour. They find that perpetrators were judged more negligent if their behaviour was preceded by exceptional circumstances. Monroe and Ysidron (2021) demonstrate that an agent's deviation from typical behaviour facilitates attributions of a particular desire and choice for this kind of behaviour. Spontaneous inference when information about mental states is absent have been argued to function as a tool to make sense of other people's past actions, promote accountability or predict future behaviour (Young & Saxe, 2009; Young & Tsoi, 2013). While the exact link between mental state inference and causal judgments in cases of single abnormal actions is yet to be shown, there is mounting evidence for mental states to play a role here, too.

2.8.3 Abnormal Objects

Deviations of 'normal behaviour' not only affect agents, but also causal attributions to inanimate objects (Cheng & Novick, 1991; Gerstenberg & Icard, 2020; Henne, Bello, et al., 2019; Knobe, 2009; Kominsky & Phillips, 2019). Gerstenberg and Icard (2020) investigate the effect of statistical normality in causal reasoning about inanimate objects in physical systems. They show their participants videos with physically realistic collisions of billiard balls, and find that participants select a statistically unlikely event as cause for an outcome in a conjunctive causal structure (see also Kirfel et al., 2020). Henne, O'Neill, et al. (2021) extend this paradigm to a variety of different inanimate objects, and show that norm-violating object behaviour affects prospective causal judgments (i.e. before the outcome occurs), independent of the perceived agency of these objects. Likewise, malfunctioning artifacts are seen as more causal for an outcome than those which function normally (Kominsky & Phillips, 2019).

Normality incorporating accounts (Icard et al., 2017; Kominsky & Phillips, 2019; Phillips et al., 2015) have the advantage of predicting causal attributions to a variety of phenomena by reference to the same mechanism, without the need to introduce additional factors such as epistemic states. Based on our findings, we argue that theories of causal attribution need to include the epistemic dimension in order to make accurate predictions about human causal agents. People do not seem to prefer abnormal agentive behaviour as causes per se, but only in combination with epistemic states (Kirfel & Lagnado, 2019; Samland & Waldmann, 2015, 2016). This finding might shed further light on the fundamental differences in our causal thinking about inanimate objects and social agents (Fausey & Boroditsky, 2007; Kelemen, 1999; Saxe, Tzelnic, & Carey, 2007; Strickland et al., 2017). As discussed earlier, counterfactual reasoning about alternatives to an action might depend on the acting agent's epistemic state. Likewise, it is possible that in case of human actions, people might prefer an agent's epistemic states as a locus of (causal) intervention (Halpern, 2016; Woodward, 2001). Given that inanimate objects do not have mental states, this additional dimension for counterfactual reasoning or intervention does not apply. As a result, including epistemic states comes at the expense of a uniform normality-based theory of causal attributions, but might sharpen our theories of causal cognition with regards to the differences between various "kinds" of causes.

2.8.4 Normal causes in Disjunctive Structures

Recent studies show that the change from a conjunctive to disjunctive causal structure flips people's preferences to a normal, typical or frequent cause (Gerstenberg &

Icard, 2020; Icard et al., 2017). If the motion detector scenario is triggered as soon as one person enters the building, people attribute more causality to the travel agent who frequently comes in at 8:45am compared to the design agent who comes in for the first time at 8:45 that day (Icard et al., 2017). The finding that the influence of normality on causal attributions is dependent on the underlying causal structure has challenged the idea of a uniform causal preference for abnormal factors. Icard et al. (2017) propose that the influence of normality is weighted differently by the sufficiency and necessity of a causal factor. Others have argued that the correspondence between the normality of an outcome and the normality of a cause is decisive for causal judgements (Harinen, 2017; Wells & Gavanski, 1989) (see Kirfel & Lagnado, 2018, for a comparison).

In light of our model, in a disjunctive causal structure, there is no link between normality of the other agent's action and the foreseeability of the outcome. The foreseeability of the outcome is high for each agent, independent of other people's actions. In this respect, the current version of our model would not predict an epistemic difference between an abnormal and normal agent. Taking into consideration the co-variation between normality of cause and outcome, however, raises new questions about the agents' epistemic states. On the one hand, it might be argued that it not only matters whether the agent expects the outcome to occur given their action, but also whether they expect their outcome to occur given their *non-action*. In a disjunctive structure, an agent will expect their action not to be necessary (or "counterfactually relevant") for the outcome if another agent acts normally or frequently – the outcome will be caused by someone else in any case. On the other hand, an agent who typically brings about a certain outcome in a disjunctive structure might know more about this outcome, be more familiar with it, or desire or intend the outcome to occur more often, than someone who has never caused it before (Alwin, 1973; D. M. Buss & Craik, 1983).

Hence, at the current stage, a variation of our Bayesian network model might be needed in order to account for increased causal attribution to typical actions in disjunctive structures. This variation would require the inference about a latent variable that is influenced by how typical, or normal an action is. Such a variation would still be compatible with the influence of epistemic or mental states, broadly considered. Under certain circumstances people are held more responsible for expected, ‘normal’ actions (Johnson & Rips, 2015). We take these and our findings to highlight the importance to monitor and control the various inferences people make about agents’ mental state and dispositions, and their potential role in attributions of causality.

2.9 Conclusion

In the aftermath of the 2015 New Taipei Water park explosion, the event organiser was found guilty of the incident and was sentenced to five years in prison (Pan, 2016, April 26). Although modern wood factories are equipped with high tech dust collection systems, employees face hefty fines if they smoke outside of designated smoking areas. A causal analysis of incidents like these two dust explosion cases lays an important foundation for legal assessment, as well as the creation of prevention measures. Both cases are structurally similar – in both, the production of dust particles together with an act of ignition lead to a dust explosion. Yet, our judgment about the primary cause differs in these cases. We take the agent who arranged the spray of combustible dust particles to be the major cause in the outdoor festival incident, but judge the agent who ignited a cigarette to have caused the explosion in the wood factory. Crucially, in both cases, the action that is selected to be the main cause is, in the respective context, the more atypical one.

In this chapter, we have argued that the typicality of actions changes how much an agent can foresee the consequences of their action. We have shown that it is this very epistemic asymmetry between a normally and abnormally acting agent that influences people’s causal judgments. Both employee and party organiser acted abnormally, but when acting, they also could have foreseen the event of a dust explosion to a greater extent. While further research is needed on this topic, the connection between action typicality and epistemic states brings us one step closer to understanding the enigmatic role of normality in causal cognition.

Chapter 3

Epistemic Interventions in Causal Reasoning

FRIAR LAWRENCE: “Unhappy fortune! By my Brotherhood,
    The letter was not nice but full of charge,
    Of dear import, and the neglecting it
    May do much danger”
    (Shakespeare, 1858, “Romeo and Juliet”)

3.1 Introduction

In the final scene of “Romeo and Juliet”, Romeo visits the tomb of Juliet who he believes to be dead, but who actually has been put into a death-like coma by a potion given by Friar Lawrence. The letter that was sent to Romeo by Friar Lawrence with the crucial information about Juliet’s faked death never reaches him because the messenger is prevented by an outbreak of the plague. Not knowing that Juliet is alive — and that his own death will later be the actual cause of Julia stabbing herself out of grief — Romeo poisons himself. Before he drinks the poison, Romeo however is puzzled by the fact that Juliet’s features still look unusually lively (“crimson in thy lips and in thy cheeks [...]// Why art thou yet so fair?”). The popularity of Shakespeare’s play lies not least in the tragedy of its ending: Romeo’s acting in ignorance about the true state of his lover actually causes her to die. If Romeo had known that Juliet was alive, she might not have died.

Bad events such as the devastating ending of “Romeo and Juliet” naturally trigger our thinking about how things could have turned out differently (Kahneman & Miller, 1986). The ability to reason ‘counterfactually’, i.e. to imagine alternative scenarios to the actual course of events, has long been argued to underpin people’s reasoning about causation (Gerstenberg et al., 2020; Halpern, 2016; Hart & Honoré, 1959/1985; Pearl, 2009). In order to determine whether something was a cause, we imagine a hypothetical scenario in which this potential causal factor is absent, and test whether the outcome is absent as well (Lewis, 2013; Woodward, 2007). Did Romeo cause Juliet’s death? According to counterfactual theories of causation, it is assumed that an agent is judged a cause to the extent that their action made a difference to the outcome (Gerstenberg et al., 2020; Gerstenberg, Halpern, & Tenenbaum, 2015; Halpern, 2016; Hitchcock & Knobe, 2009). In fact, if not for Romeo’s poisoning himself, Juliet would likely still be alive.

3.2 Aim of this Chapter

What is the most relevant change in an alternative ending scenario of “Romeo and Juliet” such that Juliet would still be alive? Intuitively, it is not (just) what Romeo did, but rather, what he knew, or more specifically, *didn’t* know. In this chapter, we want to suggest an extension of current counterfactual models of causation. In the case of ignorant causal agents, we argue that people’s counterfactual reasoning primarily targets the agent’s epistemic state – what the agent doesn’t know –, and their epistemic actions – what they could have done to know – rather than their causal action. Integrating epistemic states into causal models and counterfactual frameworks allows us to explain why people often attribute decreased causality to ignorant agents (Hilton et al., 2016; Hilton & Slugoski, 1986; Kirfel & Lagnado, 2021; Lombrozo, 2010; Samland & Waldmann, 2016). We present a novel extension to current counterfactual accounts by introducing epistemic state variables, and we test this extension by investigating people’s causal judgments in four experiments.

3.3 For They Know Not What They Do: The State of Ignorance

In the digital age where information is often just a click or a Google search away, ignorance has become an avoidable, perhaps even frowned upon state to be in. Traditionally, however, moral philosophy and the law attach mitigating circumstances to the state of ignorance, in particular to actions that arise from ignorance. The exact qualification of an act of killing ("first degree murder" vs. "negligent homicide") is to a great extent determined by how much the agent knew about the deadly consequences of their action, determining the length of the sentence (cf. "means rea", Sayre, 1932). This reduced legal culpability for ignorant actions resonates with how people generally judge responsibility for unforeseen or accidental harm. People's judgments of wrongness and moral permissibility are less sensitive to how harmful an action is, but are overwhelmingly determined by what the agents believed the consequences of their action to be (Cushman, 2008; Young & Saxe, 2011). Judgments decrease to the extent that the caused harm was unintended or unforeseen (Cushman, 2015; Margoni & Surian, 2021; Nelson-le Gall, 1985; Young & Saxe, 2008). Models of moral judgements assume agentive epistemic states to be an integral and early criterion in the process of moral decision formation (Alicke & Rose, 2012a; Goodwin, 2014; Guglielmo & Malle, 2017; Malle & Knobe, 1997). The central role of mental states also shows in the ontogeny of moral decision-making. From an early developmental stage, children start taking an agent's knowledge and intent into account in their moral reasoning about an action (Cushman, Sheketoff, Wharton, & Carey, 2013; Margoni & Surian, 2016; Piaget, 1965; Woo, Steckler, Le, & Hamlin, 2017).

3.4 Causation and Ignorance

While ignorant Romeo might not be blamed for Juliet's death, the causal role of Romeo's acting in Juliet's death seems undisputed at first glance. Recent studies in causal cognition, however, find evidence that agents' epistemic states such as knowledge or ignorance also influence people's causal judgments (Darley &

Pittman, 2003; Hilton et al., 2016; Kirfel & Lagnado, 2021; Lagnado & Channon, 2008; Lombrozo, 2010). Agents lacking knowledge (Gilbert et al., 2015) or foreseeability of the consequences of their actions (Lagnado & Channon, 2008) are judged to be less of a cause for the outcome. In causal chains, the causality of knowing agents is rated higher than those of ignorant ones (Hilton et al., 2016; Lombrozo, 2010; McClure et al., 2007). If the proximal cause is a human action, the agent is judged as more causal if the agent was aware of the causal opportunity created by prior events (Hilton et al., 2016). Likewise, as discussed in the chapter before, people's preference for abnormal actions as causes has been shown to be moderated by the agents' knowledge states about their actions (Kirfel & Lagnado, 2021; Samland et al., 2016a). We have learned already about various studies on the close connection between the development of causal reasoning in the introduction: Children produce more causal language for knowingly caused events than for unintentionally or object-caused events (Muentener & Lakusta, 2011) and attribute more causality to a human hand engaging in deliberate, goal-directed action, rather than a mere accidental movement (Leslie, 1984; Muentener & Carey, 2010).

What makes people judge a knowing agent to be more of a cause? Given the crucial role of epistemic states for moral judgments as discussed before, one obvious line of explanation is to account for these findings with reference to moral judgements. Indeed, several theories posit that the influence of agent knowledge and ignorance on people's causal judgements merely reflect their evaluations of responsibility or blame (Alicke, 2000; Alicke et al., 2012; Samland & Waldmann, 2016; Sytsma, 2019a), expressed in judgments about causation. In that sense, ignorant agents are judged as less causal because they are perceived as less blameworthy for their actions.

On the other hand, the influence of epistemic states has been argued to demonstrate something about how people perceive the causality of agents. One such proposal is that knowing and intentional agents are perceived as more "robust" causes than ignorant and unintentional ones (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020b; Lombrozo, 2010; D. Murray & Lombrozo, 2017; Phillips & Shaw, 2015). At the core of “counterfactual robustness” (D. Murray & Lombrozo, 2017) (also called “exportable dependence”) lies the idea that people assess the causality of an agent under different contingencies. The causal relationship between the outcome and a causal agent who knows and intends the outcome of their action is less sensitive to variations in background circumstances, because the agent will aim to bring about the outcome in a variety of situations (Hitchcock, 2012; Lombrozo, 2010; Woodward, 2006). Gilbert et al. (2015) show that the increased causal attribution to an agent does not necessarily hinge on the agent intending the outcome, but is given as soon as the agent knows about what kind of outcome will result from their action. In line with the idea of counterfactual robustness, Gilbert et al. (2015) show that the increased causal attribution to knowing agents is mediated by people’s reasoning about alternate possibilities. They find that in case of causal agents who can anticipate the consequences of their actions (car accident because of malfunctioning brakes), participants generate more counterfactuals about ways the outcome could have been prevented that the actor could control (e.g. informing the driver about the issue, fixing the brakes) (Spellman & Gilbert, 2014).

In sum, *if* the influence of agent epistemic states on causal judgments reveals something about people’s causal thinking, this first evidence suggests that it might do so via counterfactual reasoning. Originating in philosophy (Lewis, 2013; Pearl, 2009), counterfactual theories have become a widely used theoretical framework in order to understand people’s reasoning about causation (Gerstenberg et al., 2020; Halpern, 2016; Halpern & Hitchcock, 2015; Hitchcock & Knobe, 2009). Yet, the exact role of epistemic states in agent causality in these frameworks remains unclear. In the following, we will summarise once again the counterfactual framework of causation, and make a concrete proposal of how this framework could be reconciled with the influence of epistemic states in causal cognition.

3.5 Counterfactual Thinking and Causation

According to counterfactual theories of causation, C is a cause of E if E is counterfactually dependent on C, that is, E would not have happened in the absence of

C (Kim, 1974; Lewis, 2013; J. L. Mackie, 1974). While this framework has been originally developed as a normative theory of causation (Lewis, 1973), it has soon been adapted to address causal intuitions (Halpern & Hitchcock, 2015) and more recently been extended to capture people's actual causal attributions (Gerstenberg et al., 2020). Counterfactual dependence is assessed in terms of hypothetical interventions (Halpern, 2016; Pearl, 2009; Woodward, 2001) or mental simulations (Gerstenberg et al., 2020) over causal candidate variables, often represented in form of a *do*-operator, $do(X = x)$ (Pearl, 2009). Frameworks have cashed out the do-operator differently, e.g. setting a variable to a certain value (Pearl, 2009) or removing the causal candidate from the scene (Gerstenberg et al., 2020). Halpern (2016) extends the test for counterfactual dependence to different contingencies, i.e. non-actual possible worlds in which certain variables are set to different values. Testing for counterfactual dependence under different circumstances allows us to capture causal judgments in cases of *preemption* (Hall et al., 2003), or *overdetermination* (Gerstenberg, Halpern, & Tenenbaum, 2015; Lagnado et al., 2013).

Counterfactual models of causation are able to capture various structural aspects that influence people's causal judgments about a cause (Gerstenberg et al., 2020; Gerstenberg, Halpern, & Tenenbaum, 2015), such as causal structure, number of causes, temporal order, probabilities etc. (Gerstenberg, Goodman, et al., 2015a; Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Stephan, 2021; Henne, Kulesza, Perez, & Houcek, 2021; Icard et al., 2017; Woodward, 2011), both for inanimate causal factors as well as causal agents (Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Icard, 2020; Icard et al., 2017). However, a crucial factor that has not yet been considered in detail is the influence of epistemic states on causal judgments about agents (Hilton et al., 2016). In the case of social or agent causation, it is often assumed that the variable that is intervened on in the counterfactual scenario is the agent's action (Gerstenberg et al., 2020; Halpern, 2016; Woodward, 2011).[1] Consider as an example the case in which a doctor treats her patient with a new drug, and the patient later suffers from unexpected health prob-

[1] Kominsky and Phillips (2019) suggest that people's generation of counterfactuals include the agent's decision not to act, or to act differently.

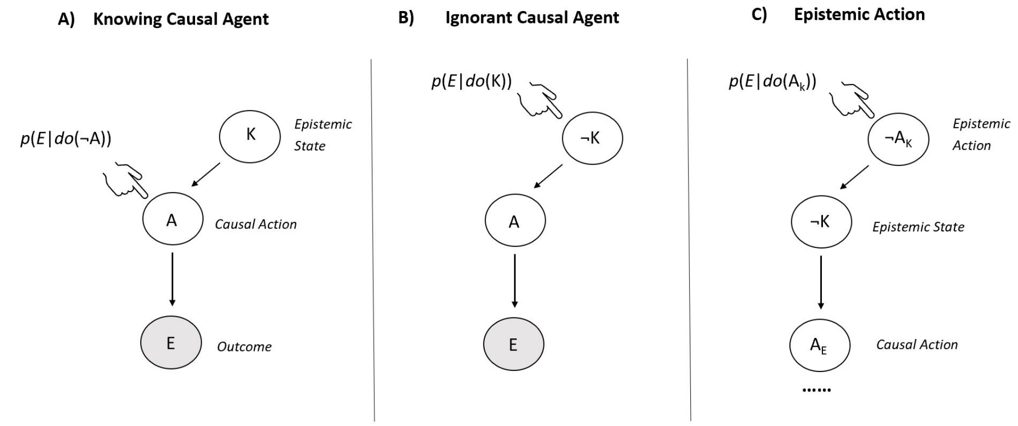


Figure 3.1: Causal Models including Epistemic States. If an agent knows about the causal outcome of their action (A), counterfactual intervention targets the agent's action A. In case of ignorant causal agents, counterfactual intervention targets the agent's epistemic state of ignorance $\neg K$ (B), or when the epistemic conditions are known, the agent's epistemic actions A_{K} that could update their state of knowledge (C).

lems. According to counterfactual theories, the doctor is a cause of the outcome to the extent that the undoing of her action would have lead to a difference in the outcome, the patient's health. Such a counterfactual dependence test, however, is insensitive to the agent's epistemic states, as it would render the doctor a cause of the health problems, irrespective of whether or not the doctor knew about the unknown side effects of the drug. While some proposals have been made on how to integrate mental states into formal frameworks of causation (Barbero, Schulz, Smets, Velázquez-Quesada, & Xie, 2020; Halpern & Hitchcock, 2015) and responsibility or blame (Chockler & Halpern, 2004; Gerstenberg, Halpern, & Tenenbaum, 2015), the general framework in its current form remains at odds with the influence of epistemic states discussed earlier (Hilton et al., 2016) – the perceived causal difference between a causal agent who is ignorant vs. knowledgeable.

3.6 A new Proposal for Counterfactual Causation: Intervening on Epistemic States

Gilbert et al. (2015)'s study suggests that the agent's knowledge state influences people's thinking about whether, and more so, *how* the outcome could have been undone: in case of knowing agents, people are able to imagine more ways how the agent could have acted such that the outcome would not have occurred. The assumption, however, that the focal counterfactual intervention targets what the agent does, i.e. the agent's actions, remains.

Our proposal aims to resolve the gap between counterfactual causal frameworks and the influence of knowledge states on causal judgements in a different manner. In causal models, social agents are usually represented by a single variable, most commonly with their causal action – i.e. by an "action" node (although see Halpern and Hitchcock (2015) for the integration of "legal" mental states into structural equation models). Extending counterfactual theories for agents' epistemic states might resolve the tension between their impact on causal judgments on the one hand and the simplified application of the *do*-operator in case of agent causation. Rather than undoing the agent's causal action A (or removing the agent from the causal scene), we argue that the primary intervention that people perform in counterfactual reasoning about ignorant agents targets the agent's epistemic state K about a certain state of the world.

Let us consider again a case in which an agent performs a certain action A, doctor Jones gives her patient a new drug, and either knows K (Figure 3.1 A) or does not know $\neg K$ that this drug has a certain side effect (Figure 3.1 B). As a result, the patient suffers from the drug's side effect E. In case of a doctor who knows about the side effect, the relevant counterfactual test concerns what would have happened had the doctor not prescribed the drug $do(\neg A)$. However, in case of an ignorant doctor (Figure 3.1 B), intuitively, this might not be the most relevant intervention people consider. Rather than undoing the doctor's action, people might want to primarily change her epistemic state from ignorance to knowledge about the side effects of the drug, $do(K)$. When people are faced with causal outcomes caused by an agent, we hypothesise that they build richer causal models including not only the agent's causal actions, but also their mental states, in order to determine causality. Drawing on a causal model framework, we argue that in counterfactual reasoning, people represent an agent's mental states, and that counterfactual intervention, in

Pearl (2009)'s terms, the $do()$ operator can target epistemic variables. We see two broad arguments for this claim.

3.6.1 Why Epistemic Intervention?

3.6.1.1 Changing Actions via Knowledge

On the one hand, epistemic states, and mental states more broadly, often act as preconditions for the change of an action (Gibbons, 2001; Hawthorne & Stanley, 2008; Webb & Sheeran, 2006). There is a variety of evidence showing that people do not intervene on any variable but apply counterfactual intervention selectively, depending of the perceived "mutability" of a target variable (Dehghani, Iliev, & Kaufmann, 2007, 2012; McGill & Tenbrunsel, 2000) (also referred to as a "availability" of a particular counterfactual (Bear, Bensinger, Jara-Ettinger, Knobe, & Cushman, 2020; Icard, 2016; Kahneman, 2014; Phillips, Morris, & Cushman, 2019). Walsh and Byrne (2007b) as well as Bonnefon, Zhang, and Deng (2007) show that mental states such as an agent's reasons for an action influence to what extent people actually imagine an undoing of the action in counterfactual reasoning (see also Bonnefon, 2007; Juhos et al., 2015). Likewise, in the most basic form our argument relies on the assumption that for an agent to act differently, the epistemic variable about certain properties of that action or the world must be set to a certain value. Changing the epistemic state from e.g. 'not-knowing' to 'knowing' about a fact related to the action facilitates changing the action variable. A test for counterfactual dependence of an outcome on a causal agent who lacks knowledge about a certain state of the world will hence require the hypothetical intervention on the agent's epistemic state. In more intuitive terms, only in a world in which the doctor knows about the side effect of the drug can they decide not to prescribe it, or will be more likely not to do so. We can summarise these ideas in two counterfactual conditionals that we take to approximate people's reasoning about 1) *Knowing Causal Agents* and 2) *Ignorant Causal Agents*:

1. *Knowing Causal Agent* If S had not A-ed, E would be different.

2. *Ignorant Causal Agent* If S had known that p [and not A-ed], E would be different.

How does the hypothetical intervention on an agent's epistemic state affect the agent's action? Changing an agent's epistemic state from ignorance to knowledge about the consequences of their action might also bring about a change in the agent's action, but might not automatically undo the action. Will Dr Jones refrain from prescribing the drug if she knows about the side effect? The exact probability of a change in action given some state of knowledge will vary, depending on further assumptions such as the agent's general preferences, motivations for the action, etc (Weinstein, 1972). These assumptions will depend on both context as well as individual priors. However, assessing the likelihood of a difference in the outcome is now dependent on two things: the probability of the agent acting (or not acting) given knowledge, and the probability of the outcome given the agent's action (or non-action). In contrast to a case in which the intervention directly targets the causal action, intervening on a prior (epistemic) variable increases the uncertainty about whether outcome will turn out different. We think that this weakened counterfactual dependence between targeted variable and outcome in case of epistemic interventions might be one reason for the weaker perceived causal strength of ignorant agents.

3.6.1.2 Future Causality

In addition, we argue there exists an additional motivation for the sketched proposal here. Recent work in causal cognition has increasingly highlighted the crucial role of causal judgements in identifying targets for future intervention (N. Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017; N. R. Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Ferrante, Girotto, Stragà, & Walsh, 2013; Gerstenberg & Icard, 2020; Hitchcock, 2012). In this sense, causal judgments have a dedicated *forward-looking* function. They single out those factors that would have made a difference in the actual scenario, but that will also cause the outcome in future or similar scenarios (Grinfeld et al., 2020b; Lombrozo, 2010). Mental states make for a suitable target of "intervention" on behaviour and action outcomes, as numerous psychological studies demonstrate (Dolan, Elliott, Metcalfe, & Vlaev, 2012; Dolan, Hallsworth, et al., 2012; Kerwer, Rosman, Wedderhoff, & Chasiotis, 2021; Murphy & Mason, 2006). Only an agent who possesses knowledge about the consequences of their action can and will effectively adapt their behaviour accordingly (Ajzen, 1985). Independent of whether the target of intervention is to ensure the outcome to happen in future or for it to be prevented, it is crucial that the agent has relevant knowledge about it. This forward-looking function might also explain why people assign less causality to agents bringing about accidental positive outcomes (Guglielmo & Malle, 2019; Lagnado & Channon, 2008; Malle, Guglielmo, & Monroe, 2014a). Only an agent who knows about the positive outcomes of their action can ensure to bring them about again, or is more likely to do so than an ignorant one. Intervening on the agent's epistemic state about a positive outcome increases the agent's future causality for a good outcome (Gerstenberg et al., 2020; Grinfeld et al., 2020b; Icard et al., 2017). Intervening on epistemic states hence also marks an intervention that makes the agent a robust cause for the targeted outcome state.

3.6.1.3 Epistemic Actions

Once we acknowledge the role of epistemic states in people's causal models, this also raises the question *how* this epistemic state change could have been obtained. Could the doctor have known that the drug causes side effects, and if so, what could the doctor have done to know? When the epistemic context of an ignorant causal agent is known, we argue that people not only represent the agent's epistemic states, but also their *epistemic actions* (Kirsh & Maglio, 1994; S. Miller, 2018) — actions the agent could have done in order to gain the relevant piece of information that would led to knowledge (Figure 3.1 C). For example, imagine that information about side effects usually comes in a package leaflet with the drug, but that the doctor fails to read the leaflet. Ignorant about the side effect, she prescribes the drug and the patient suffers from side effects. Rather than unspecifically intervening on the agent's epistemic states, the relevant intervention here seems to target the agent's epistemic (non)-action, the not-reading of the leaflet. We can express this intuition by distinguishing between epistemic actions A_K that are connected to the epistemic state K, and causal actions A_E that are connected to the outcome E (Figure 3.1 C).

3. *Epistemic Action* If S had A_K-ed [S would know that p [S would not have A_E-ed]], E would not have happened.

The causal representation of epistemic actions and epistemic context can flexibly capture what needs to happen in order for an agent to acquire knowledge. On the one hand, a causal model that is enriched by variables encoding an agent's epistemic condition can include the *number of epistemic actions* that are necessary to change the agent's epistemic state. On the other hand, such a model can also capture the *epistemic contingencies*, for example background variables like the availability of information that determines whether an epistemic action will successfully change the epistemic state or not. Counterfactual theories of causation consider whether counterfactual dependence is obtained in the actual world, but also under different 'contingencies', i.e. when variables are set to different values (Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2014; Halpern, 2016). Assuming that people represent epistemic variables, this in consequence applies to epistemic states as well: That is, testing not only what the agent could have known in the actual world, but also what they could have known under different circumstances.

3.7 Hypotheses

In this chapter, we aim to empirically test the outlined proposal. More precisely, we aim to test whether people's causal judgments are sensitive not only to the agent's epistemic state (1. & 2.), but also to the agent's epistemic actions (3.). We hypothesise that people's causal judgements about ignorant agents reflect their counterfactual intervention on epistemic variables in a causal model. In case of ignorant causation, people will change the agent's epistemic states and/or the actions the agent could have performed to acquire knowledge. Hence, we hypothesise the following:

(i) **Hypothesis 1**

  (a) *Causal Judgment:* Ignorant agents are judged as less causal than knowing agents.

  (b) *Counterfactual Reasoning:* If the causal agent is ignorant, people intervene on epistemic states, rather than causal actions.

(ii) **Hypothesis 2**

  (a) *Causal Judgment* Ignorant agents who could have changed their epistemic states are judged causal to the extent that they could have acquired knowledge.

  (b) *Counterfactual Reasoning* If the causal agent is ignorant, people intervene on the agent's *epistemic actions*, rather than causal actions.

Hypothesis 1 aims to test more generally whether people aim to intervene on epistemic states. Comparing a knowledgeable vs. an ignorant causal agent, we predict that the latter is judged less causal, and that people's imagined counterfactuals will target the agent's epistemic state (*Experiment 1: Knowing vs. Ignorant Agents*). Hypotheses 2 makes predictions for cases in which a causal agent is ignorant, but could — by their own epistemic actions — have changed their state of ignorance and acquired knowledge. We will test Hypothesis 2 for three different cases of epistemic action conditions. In the most basic case, an agent is ignorant, and either could or could not have acquired knowledge by their own action (*Experiment 2: Externally vs. Self-Caused Ignorance*). In such a case, we predict that the agent who could have changed their epistemic state will be judged as more causal, and that people will intervene on this agent's epistemic (non-)action. A different scenario which tests this hypothesis is a case in which two ignorant agents both could have acquired knowledge, but differ in how many actions it would have taken them to acquire knowledge (*Experiment 3: Number of Epistemic Actions*). Here, the agent for whom it would have been easier to acquire knowledge should be judged more causal for the outcome. Finally, we turn back to the idea that counterfactual dependence is assessed under different contingencies (*Experiment 4: Epistemic Actions under Contingencies*). We predict that an agent whose epistemic action did not lead them to acquire knowledge, but would have led them to acquire knowledge under different circumstances, is judged less causal than one who would remain ignorant in both actual and possible world.

3.7.1 Individual Causal Models of the World

The central assumption underlying our hypothesis is that people's causal judgments reflect counterfactual interventions on the causal models they built of a causal scenario (Gerstenberg et al., 2020; Gerstenberg & Lagnado, 2014; Halpern, 2008). However, people's mental representation of causal variables and causal structure, and in particular their reasoning about what could have gone different might vary individually (Kasimatis & Wells, 1995; Roese & Olson, 2014; Rottman, Gentner, & Goldwater, 2012). Consider again that in Shakespeare's drama, Romeo did not know that Juliet is alive because the letter with the relevant information couldn't be delivered. But could Romeo have undertaken any alternative actions in order to find out about her true state? People might have different assumptions about whether and what Romeo could have done to acquire the relevant knowledge that potentially would have led him to act differently.[2] In line with our argument, such assumptions about the possibility of epistemic actions that could have led Romeo to acquire knowledge will influence people's causal judgments about him. We predict that, in addition to the epistemic context of the scenario, people's causal judgments will differ by their subjective beliefs about whether and how easily the agent could have changed their epistemic state.

3.7.2 Blameworthiness for Ignorance

Some moral philosophers argue that an agent's blameworthiness for an unknown consequence of their action derives from their blameworthiness for their state of ignorance (Rosen, 2004; Wieland & Robichaud, 2017; Zimmerman, 1997). Blame-

[2] As one Reddit user comments, "Romeo and Juliet would still be alive if he had checked her pulse."("invertedparadoxxx", 2019))

worthiness for ignorance is given in case of a “benighting” omission (or action) that the agent was able to control, that is “all-things-considered wrong”, and that causes the agents to lack the relevant knowledge about the outcome of their actions (Smith, 1983). According to theories of “derivative blameworthines” for ignorance, the predictions of Hypothesis 2 should apply to judgments about blameworthiness for ignorance. An agent will be held blameworthy for their ignorance (and hence for the outcome of their ignorant action) if they could have performed some kind of epistemic action that would have caused them to possess knowledge. In addition to judgments about causation, we will hence also assess people’s judgments about blame for ignorance in the different epistemic conditions sketched above. Causality has been argued to be one of the major building blocks for judgments of responsibility and guilt, but the debate about how causation and blame relate, especially in people’s cognition, is ongoing (Alicke, 2000; Knobe, 2009; Samland & Waldmann, 2016; Shaver & Drown, 1986). We will return to the discussion about the relationship between causality and blame for ignorant causal agents in the General Discussion, and sketch how these fit into the counterfactual picture that we suggest.

3.8 Experiment 1

In Experiment 1, we aim to investigate people’s causal judgments and counterfactual reasoning about a knowing (Figure 3.1 A) vs ignorant causal agent (Figure 3.1 B), i.e. an agent who is either aware or unaware of the causal consequences of their action.

3.8.1 Participants and Design

We recruited 145 participants on Amazon Mechanical Turk. 23 participants were excluded for failing one or more of the four comprehension check questions, and one participant was excluded for providing non-sensical counterfactual responses, leaving a final sample size of $N = 121$ (M_{age} = 38.42, SD_{age} = 11.15, N_{female} = 40). We adopted a 2 knowledge (knowledge vs. no knowledge) × 3 scenario (“hospital” vs. “garden” vs “bakery”) design. ‘Knowledge’ was manipulated within participants and ‘scenario’ was manipulated between participants.

3.8.2 Material and Procedure

Participants read both the 'knowledge' as well as the 'no knowledge' condition of one of the three scenarios ("hospital", "garden", "bakery") in randomised order. All three scenarios follow the same content structure: As part of their work, an agent usually applies a certain product ("medical drug", "fertilizer", "baking flour"). A newly acquired product is of the same quality, but has potentially harmful properties or consequences.

> (Vignette "Hospital")
> "Dr Jones works as a doctor in a local hospital. Dr Jones often administers her patients the blood-thinning drug "Heparine" in order to prevent thrombosis and blood clots. Normally, blood-thinning drugs do not cause any side effects with certain blood types.
>
> The hospital has recently started to order an additional blood-thinning drug, 'Afibo', that is cheaper than 'Heparine'. 'Afibo' is as effective as 'Heparine', but has one side effect. It causes mild leg cramps in patients with blood type 'AB-positive'. "

Depending on the 'knowledge' condition, the middle part of the vignette manipulated whether the agent possesses relevant knowledge about the harmful properties of the item.

> *Knowledge* "Although the drug 'Afibo' has only recently been ordered, Dr Jones knows that this drug causes mild leg cramps in patients with blood type 'B-negative'. "
>
> *No Knowledge* "Because the drug 'Afibo' has only recently been ordered, Dr Jones does not know that this drug causes mild leg cramps in patients with blood type 'AB-positive'. "

After reading the first part of the vignette, participants had to answer two comprehension check questions: First, a question about the outcome, 1) "The new blood-thinning drug 'Afibo'..." i) "... causes mild leg cramps in patients with blood type 'AB-positive', ii) "... causes sore throat in patients with blood type 'AB-positive'." And second, question about the agent's knowledge state: 2) "Dr Jones ..." i) " ... knows that 'Afibo' causes side effects in patients with blood type 'AB-positive'." , ii) "... does not know that 'Afibo' causes side effects in patients with blood type 'AB-positive'." The final part of the vignette describes the agent's use of the item, resulting in harmful consequences. Participants then proceeded to the last part of the vignette.

> "One day, Dr Jones is treating a patient. Checking the patient's medical record, Dr Jones sees that the patient has blood type 'AB-positive'. Dr Jones knows [does not know] that the new blood thinner"Afib" causes mild leg cramps in people with blood type 'AB-positive'.
>
> Dr Jones administers "Afibo" to that patient. The drug helps to prevent the patient's onset of thrombosis, but the patient also suffers from mild leg cramps."

3.8.2.1 Causal Rating Question

After the final part of the vignette, participants had to answer a causal rating question, and generate an counterfactual alternative in an open-text response. The causal rating question asked participants to what extent they agree with the statement "Dr Jones [agent] caused the patient's leg cramps [outcome]" on a 7-point Likert scales (1-'strongly disagree', 7-'strongly agree').

3.8.2.2 Counterfactual Response Question

For the counterfactual response question, participants were instructed to write down what could have gone differently so that the patient would not have suffered mild leg cramps. For orientation, they were provided with the example sentence "If ___, the patient would not have suffered leg cramps [effect absent]". Participants were informed that their response does not need to fit into exactly into the format of the example sentence and can be as long as needed. Participants wrote their answer into an open text-box with unlimited character length. This open-text counterfactual question allowed us to elicit the individual point of intervention in people's

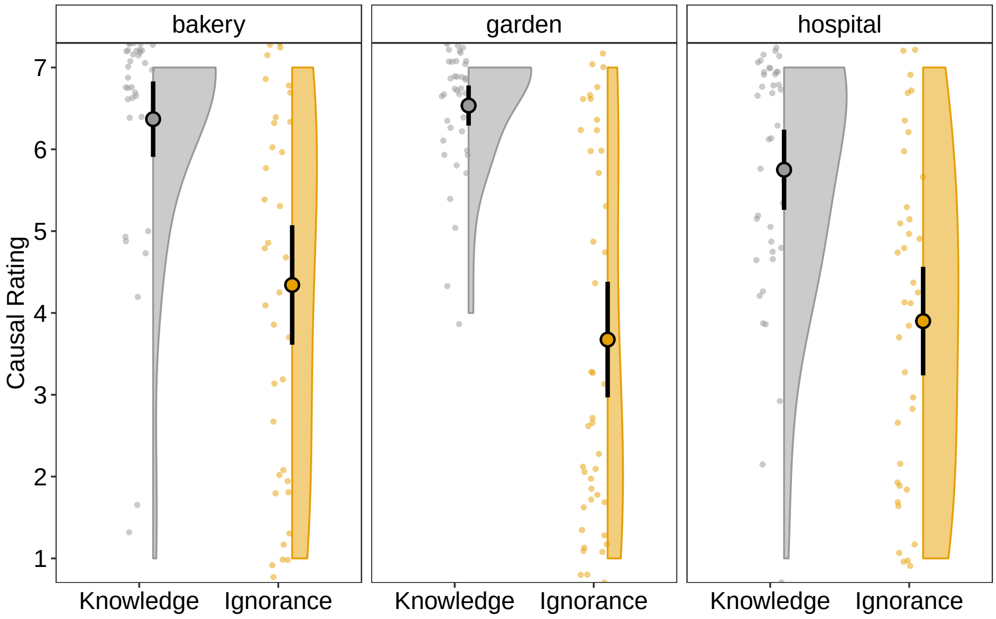


Figure 3.2: Experiment 1: Causal Ratings. Big dots are group means. Error bars depict 95% Confidence Intervals. Coloured backgrounds represent the probability distribution of the data, grey dots are individual participants' judgments jittered for visibility.

imagined alternative scenarios.

At the end of the experiment, participants provided demographic information and were thanked for their participation in the study.[3]

3.8.3 Analysis of Data

The central manipulation in our experiment — the knowledge manipulation — was employed as a within-participants variable. In Experiment 1 (as well as the following experiments) we analysed participants responses to the different knowledge manipulations both as *within-subject* as well as *between-subject* contrast.

We analysed the within-subject effect of knowledge and scenario on participants' causal ratings as within contrasts by fitting linear mixed effects models to the data using the *lmer* (Bates, Mächler, Bolker, & Walker, 2014) and the *afex* package (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2020). The model included 'scenario' and 'knowledge' as fixed effects and participants as random intercepts.

[3] All materials of the experiments, data and analysis code can be found here: `https://github.com/LaraKirfel/EpistemicInterventions`

We analysed participants' open text counterfactual responses by using multinomial regression with the *VGAM* package (Yee et al., 2010) to model the relationship between knowledge and response type membership.

We also analysed these effects as between contrasts. For this, we used only the data from the first scenario, i.e. either the 'knowledge' or 'ignorance' condition, that participants saw in our experiments. We did so using a series of linear regression models with the *lme4* (Bates, Sarkar, Bates, & Matrix, 2007) and *car* package. In the results section, we will report the test statistics for both within- and between-subject effects, with between-subject test statistics in brackets. In order to keep the results section concise, descriptives of only the within-subject effects are reported in text, and depicted in the figures. The descriptive statistics for the between-subject effects can be found in Appendix B.

3.8.4 Results

3.8.4.1 Causal Rating

Including the factor knowledge into a model provided a better fit for the data than a model without it, $\chi^2(1) = 92.52$; $p < .001$ [*between-subject contrast*: $F(1) = 35.19$; $p < .001$]. People's causal ratings were lower (b = -2.25, *SE* = .19, t = 11.58) when the agent was ignorant (*M* = 3.97, *SD* = 2.20, 95% CI [3.57, 4.35]) compared to a knowing agent (*M* = 6.22, *SD* = 1.30 95%, CI [5.99, 6.46]) (see Figure 3.2). Adding scenario $\chi^2(2) = 2.66$, $p = .27$, [*between-subject*: $F(2) = 11.7$; $p = .14$] and an interaction term with knowledge ($\chi^2(2) = 5.37$; $p = .07$) did not provide a significantly better fit to the data [but in the *between-contrast*, $F(1) = 92.5$; $p <$ .001, see Appendix B].

3.8.4.2 Counterfactual Responses

Based on participants' free text responses, we developed a coding rubric for the self-generated counterfactual responses. The coding rubric had four categories. Participants' responses were coded by the first author and a research assistant. Inconsistent codes were resolved by discussion (Inter-rater agreement: 94%).

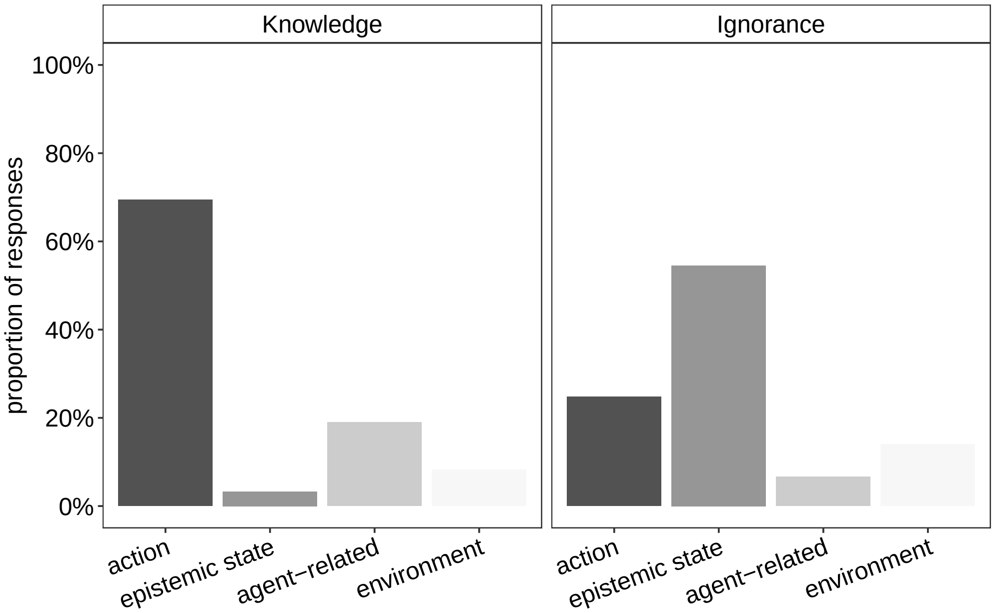


Figure 3.3: Experiment 1: Counterfactual responses. Proportion of choice of four counterfactual response categories (“action“, “epistemic state”, “agent-related”, “environment”)

- **“Action”**: The first category “Action” (*N* = 114) covered all responses that described a change of *just* the agent’s causal action that led to the outcome, i.e. the use of the fertilizer/drug/baking flour. Responses in the “Action” category would either describe direct undoing of the agent’s action (*“If Dr Smith had not administered Corus to the patient”*), or imply undoing the action by suggesting an alternative action (*“If the doctor chose Sanguine...”*).

- **“Epistemic State Change”:** The second category “Epistemic state” (*N* = 70) covered all responses that described a change in the agent’s epistemic states. The coding category included any states of knowledge, belief or expectation about the consequences of the action/item used. Responses in this category either described a direct change in the focal agent’s epistemic state (*“If Dr. Jones had known about the side effects of Afibo...”*), an epistemic change caused by an action of the agent (*“If Dr Jones had researched the side effects of Afibo before administering it...”*) or an epistemic change caused by someone else (*“If the bakery manager would have informed Anne about* *buying Homestead flour and its possible traces of hazelnut..."*).

Remaining answers did not show a specific theme, so we clustered the answers around two broad categories.

- **"Agent-related"**: The third category included any changes related to the focal agent, "Agent-related" ($N = 31$). Answers included an additional action by the agent that could have prevented the outcome (but that did not include undoing the action of giving the drug) (*"If she had warned others of the contents [of the flour]"*, prior actions by the agent that could have prevent the problem in the first place (*"If Anne had discussed not making any changes with her staff without her knowledge,...."*), but also changes in the agent's character traits (*"If Alex was truly committed to his the rose...", "If Alex' greed didn't get in his way.."*).
- **"Environment"**: The fourth category "Environment" (N = 27) included all kinds of changes that did not relate to the agent. This category included answers suggesting a change in the affected object or person (*"If the patient did not take 'Corus'...*), an action or epistemic state change by a third party (*("If the bakery manager did not order this kind of flour")*, or modifications in the item used (*"If the product didn't contain walnuts..."*).

3.8.4.3 Analysis

A multinomial logistic regression was performed to model the relationship between the knowledge condition and the type of counterfactual response ("action", "epistemic state", "agent-related", "environment"), with "environment" as reference category. Addition of the knowledge predictor to a model that contained only the intercept significantly improved the fit between model and data, $\chi^2(3)$ = 102.42; $p < .001$, R^2= .17 [*between-subject contrast*: $\chi^2(3)$ = 52.48; $p < .001$, R^2= .17].

When the agent's epistemic state changes from knowledge to ignorance, people are less likely to imagine a counterfactual change that concerns the agent's action (69% vs. 25%) (b = -1.10, OR = .33, SE = .32, z = -3.45, $p < .001$) (see Figure 3.3). In contrast, when agents are ignorant rather than knowing, people are more likely to imagine a change in the agent's epistemic state (55% vs. 3%) (b = 1.61, OR = 4.99, SE = .46, z = 3.49, $p < .001$). Finally, people are also less likely to imagine a change related to the agent in general when the agent is ignorant (7% vs. 19%) (b = -1.12, OR = .32, SE = .40, z = -2.77, $p < .01$) (see Figure 3.3).

3.8.5 Discussion

The first experiment replicated previous findings demonstrating the influence of agent epistemic states on people's causal attributions (Gilbert et al., 2015; Lagnado & Channon, 2008; Lombrozo, 2010): Ignorant agents are perceived as less causal for an outcome than knowledgeable agents. At the same time, the agent's epistemic state also shifts the target of an imagined counterfactual intervention. In case of ignorance, people are less likely to refer to a change in the agent's causal action, but prefer to imagine a change in the agent's epistemic state, and more precisely, a change from ignorance to knowledge (Hypothesis 1). We showed that this pattern holds as a within-contrast, comparing participants' responses across both conditions, as well as between-contrast, comparing participants' responses to the first scenario that they saw.

Experiment 1 provides first evidence for our hypothesis that people naturally represent and refer to agent's epistemic states when engaging in counterfactual reasoning. The fact that an agent who knows about the harmful consequences of their action still proceeds to perform this action raises the question about what further inferences people made about the agent in the "knowledge" condition of this experiment (Gerstenberg et al., 2018a; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Siegel, Crockett, & Dolan, 2017). While additional inferences about the agent's character are possible factors that might have influenced people's causal judgments, it is noteworthy however that people's counterfactual responses overwhelmingly referred to the undoing of the knowing agent's action, rather than

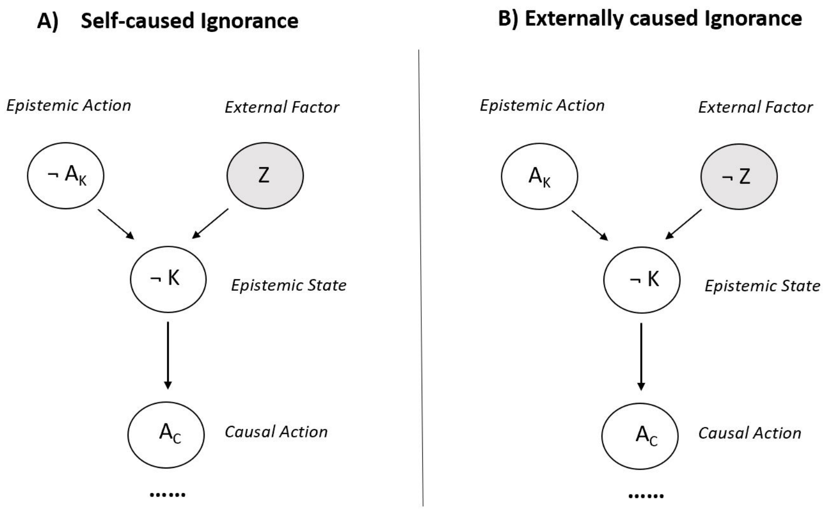


Figure 3.4: Experimental Conditions of Experiment 2. In the "Self-caused Ignorance" condition (A), the agent does not perform the epistemic action that would update their knowledge, although the external condition are given, i.e. the information is available. In the "Externally Caused Ignorance" condition (B), the agent aims to inform themselves, but the information is not available.

a change in the agent's character traits or dispositions ("If only Dr Smith would not have been so malicious..."). Further inferences about additional mental or dispositional states hence leave the focus on an action intervention in counterfactual reasoning about knowing agents unchanged. In the General Discussion, we will return the general role of blame and how it fits in the account we propose here.

In the scenarios of Experiment 1, the exact reasons for the agent's ignorance about the consequences of their action were underspecified. In the "ignorance" condition, our experimental scenarios leave open whether and to what extent the agent could have changed acquired the relevant knowledge. Our Hypothesis 2 predicts that the epistemic conditions of an ignorant causal agents matter for people's causal assessments. We were therefore interested if the conditions under which an agent's ignorance came about also influence how causal the agent is perceived, as well the kind of counterfactuals people imagine. This question was the aim of Experiment 2.

3.9 Experiment 2

In the second experiment, we aimed to assess judgments about ignorant causal agents whose ignorance was either self- or externally caused (see Figure 3.4). More specifically, in the "self-caused ignorance" condition, the external conditions for information acquisition are given, but the agent does not perform the necessary epistemic action in order to acquire the information (see Figure 3.4 A). In contrast, in the "externally caused ignorance" condition, the agent aims to acquire knowledge, but the necessary external conditions for obtaining the information are not given (see Figure 3.4 B).

3.9.1 Participants and Design

We recruited 179 participants on Amazon Turk. 27 participants were excluded for not answering all eight comprehension check questions correctly, and two participants were excluded for providing a nonsensical counterfactual responses. The final sample consisted of 150 participants (M_{age} = 37.78, SD_{age} = 11.67, N_{female} = 59). We adopted a 2 ignorance (self-caused vs. externally caused) × 3 scenario ("hospital" vs. "garden" vs "bakery") design. 'Ignorance' was manipulated within participants and 'scenario' was manipulated between participants.

3.9.2 Material

The main story was the same as in Experiment 1, but this time agents were ignorant about the consequences of their action in both conditions. However, what differed was how their state of ignorance was brought about. In this vignette, an e-mail that contains the relevant information about the harmful properties of an item is sent to the agent.

> "The pharmacy manager has sent an e-mail with information about the new blood-thinning drug "Afibo" to all doctors in the hospital. The e-mail contains the information that the drug will cause mild leg cramps in patients with blood type 'AB-positive"'.

In the *"externally caused ignorance"* condition, this e-mail is however deleted due to a technical default.

> "The e-mail service provider of Dr Jones has recently upgraded its e-mail service. Because of an undetected bug in the upgrade, the e-mail filter settings for spam content have changed. Dr Jones checked her inbox, but she did not see the e-mail of the pharmacy manager because it was erroneously marked as spam and automatically deleted from the account."

In the *"self-caused ignorance"* condition, the agent does not obtain the information because they fail to read the e-mail.

> "Dr Jones checked her inbox and saw the e-mail of the pharmacy manager, but did not read it."

In both conditions, the scenarios ends with the agent applying the relevant item, unknowing about the harmful properties of the item. As a result, a bad effect obtains.

3.9.2.1 Causal Rating and Counterfactual Question

Causal and Counterfactual Question were asked as in Experiment 1. Agreement with the causal statement was assessed on a 7-point Likert scales (1-'strongly disagree', 7-'strongly agree') - "Dr Jones [agent] caused the patient's leg cramps [outcome]" -, and counterfactual responses were given in an open text box: "If ___, the patient would not have suffered leg cramps [effect absent]".

3.9.2.2 Knowledge and Blame Rating

In addition to the perceived possibility of knowledge, we also wanted to assess people's judgments about the agent's blameworthiness for their ignorance. Hence, we added two questions asking for people's modal judgment about the agent's epistemic state, and for the agent's blameworthiness for their ignorance. . Participants had to indicate their agreement with the modal statement "Dr Jones [agent] could have known that 'Afibo' causes leg cramps [effect]" on a 7-point Likert scale (1-'strongly disagree', 7-'strongly agree'). Finally, participants had to answer the question "How blameworthy is Dr Jones [agent] for not knowing that 'Afibo' causes leg cramps [effect]?" on 7-point agreement scale (1-'Not at all', 7-'Completely').

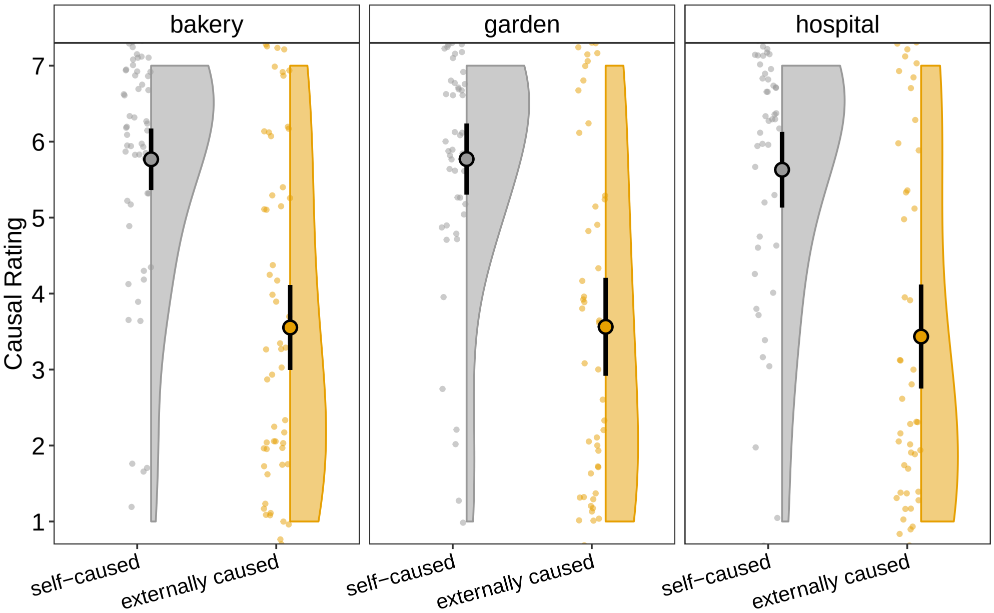


Figure 3.5: Experiment 2: Causal Ratings. Big dots are group means. Error bars depict 95% Confidence Intervals. Coloured backgrounds represent the probability distribution of the data, grey dots are individual participants' judgments jittered for visibility.

3.9.3 Results

3.9.3.1 Causal Rating

Likelihood ratio test indicated that type of ignorance was a significant factor in predicting participant's causal responses, $\chi^2(1) = 108.54$; $p < .001$ [*between-contrast*: $F(1) = 23.85$; $p < .001$]. People's causal ratings decreased (b = -2.21, SE = .18, t = -12.59) when the agent's ignorance was caused externally (M = 3.52, SD = 2.19, 95% CI [3.17, 3.87]) rather than by choice (M =5.73, SD = 1.59 95%, CI [5.48, 5.98]) (see Figure 3.5). There was no significant effect of scenario (p = .90) [*between-contrast*: p = .72] and no interaction between ignorance and scenario (p = .99) [*between-contrast*: p = .44].

3.9.3.2 Counterfactual Reasoning

Based on participants' free text responses, we developed a coding rubric for the counterfactual responses. We excluded the responses from eight participants who indicated that the agent in the "externally caused ignorance" condition could have

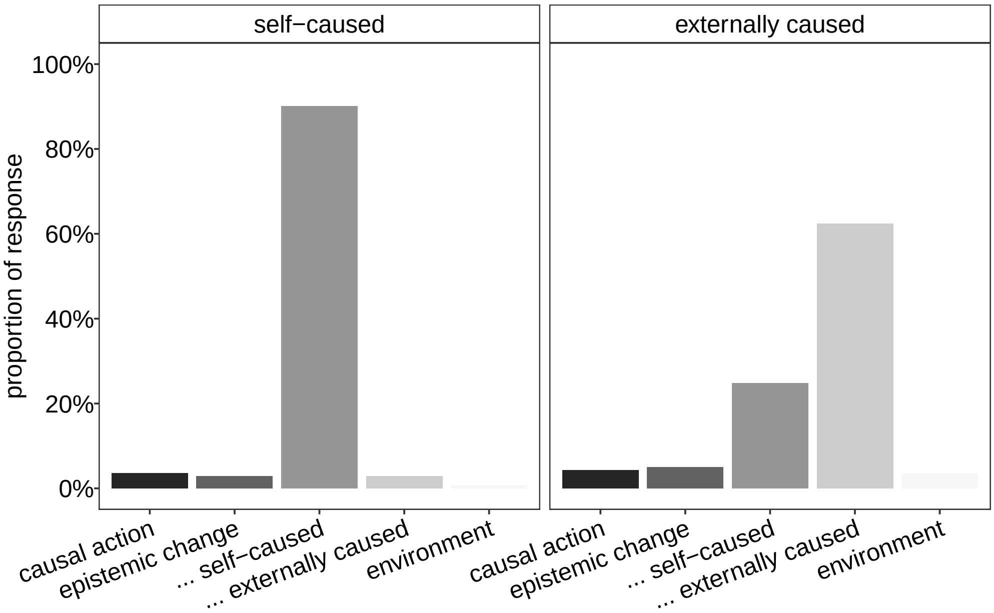


Figure 3.6: Experiment 2: Counterfactual responses. Proportion of responses in the four counterfactual response categories (“action“, “epistemic state change”, “self-caused epistemic change”, “externally caused epistemic change”, “epistemic change by other”, “environment”)

looked into the spam-folder and read the e-mail, signalling a misunderstanding of the scenario. Inter-rater agreement was at 91%.

- **“Action”**: The first category “Action” (N = 11) covered all responses that described a change in the agent’s action that caused the outcome, either by undoing the actual action (*“If Alex did not fertilize [...] the Bourbon roses with the fertilizer ”Splendor”...”*) or by an alternative action (*“If the patient would have been prescribed Heparine...”*).

Many responses suggested a direct or indirect change of the agent’s epistemic state, but differed in how this knowledge change is brought about. We created three broad categories that roughly capture the various ways people imagined the agent’s change in knowledge state: i) direct (without further specification), ii) by an action of the agent, iii) by an action or cause that is unrelated to the agent.

- **"Direct epistemic change"** (*N* = 11) referred to responses that suggested a direct change of the agent's knowledge about the item without specifying how (*"If Sandra had known about the walnuts..."*).' We also included in this category 'Epistemic state change about technical failure" (*N* = 1). The response in this category indicated a change in the agent's knowledge about the technical default in the e-mail system (*"If the doctor had known about the bug in the email system..."*).

- **"Self-caused epistemic change"** (*N* = 162) included all types of epistemic state changes of which the agent was the primary cause. On the one hand, this included the sub-category "... by reading the e-mail", the aqcuistion of knowledge by reading the relevant e-mail (*"If Bob had read his email..."*). Secondly, this category included additional actions of the agent "... by an additional action by the causal agent.". This category referred to responses indicating the focal agent performing an action (independent of the e-mail) that leads to the relevant knowledge (*"If the doctor had done their own follow up research"*, '*'If Sandra would have read the label on the new flour or asked the bakery manager about the new flour and if it was ok to use..."*).

- **"Externally caused epistemic state change"** (*N* = 92) included changes in the agent's epistemic that were not primarily caused by an action of the agent themselves. The sub-category " ... by an e-mail that's made accessible (due to a technical fix, etc.)" referred to responses naming a variety of causes that made the inaccessible e-mail accessible (*"If the spam filter didn't mark the email as spam..."*, *"If the bug does not happen while the upgrade from e-mail service provider of Sandra ..."*). Responses in the sub-category "... by being informed by a third-party agent" referred to the causal agent being informed by a third party, or an additional/different action by a third party agent (*"Dr. Jones would have been informed in a letter or face to face format"*, *"The email sender should have confirmed Sandra received the information or should have called here and informed her instead of emailing her."*).

- **“Environment”**: Finally, the category “Environment” (N = 6) comprises changes in the environment or setting that do not directly affect the agent’s causal action or epistemic state (*“If the company was not cutting corners and using cheap flour...”*).

A multinomial logistic regression was performed to model the relationship between the ignorance condition and the type of counterfactual response (“action”, “direct epistemic state change”,“epistemic change by agent” “epistemic change by other”, “environment”), with “action” as reference category. Addition of the ignorance predictor to a model that contained only the intercept significantly improved the fit between model and data, $\chi^2(4) = 153.96$; $p < .001$, $R^2 = .27$ [*between subject contrast*: $\chi^2(4) = 53.87$; $p < .001$, $R^2 = .17$].

Changing the epistemic condition of ignorance from self-caused to externally caused is associated with a decrease in the relative log odds of indicating a self-caused epistemic change (b = -1.04, OR = .35, SE = .45, z = -2.32, p = .02). People were less likely to imagine a self-caused epistemic change in the externally vs. self caused ignorance condition (25% vs. 90%) (see Figure 3.6)

In contrast, the change from self-caused to externally caused ignorance increases the relative log odds to indicate an externally caused epistemic change over an action change (b = 2.06, OR = 7.82, SE = .56, z = 3.67, $p < .001$). People are more likely to imagine an epistemic state change that is caused by external or other factors in the external vs. self-caused condition (62% vs. 3%).

3.9.3.3 Subgroup Analysis: Causal Rating by Counterfactual Response

In the externally caused ignorance condition, there was a substantial proportion of people who indicated that the agent could have obtained knowledge by an action of their own (25%). We were interested in analysing people’s causal rating in dependence of what kind of counterfactual response they gave. That is, we wanted to investigate whether causal ratings generally differed between people who imagined

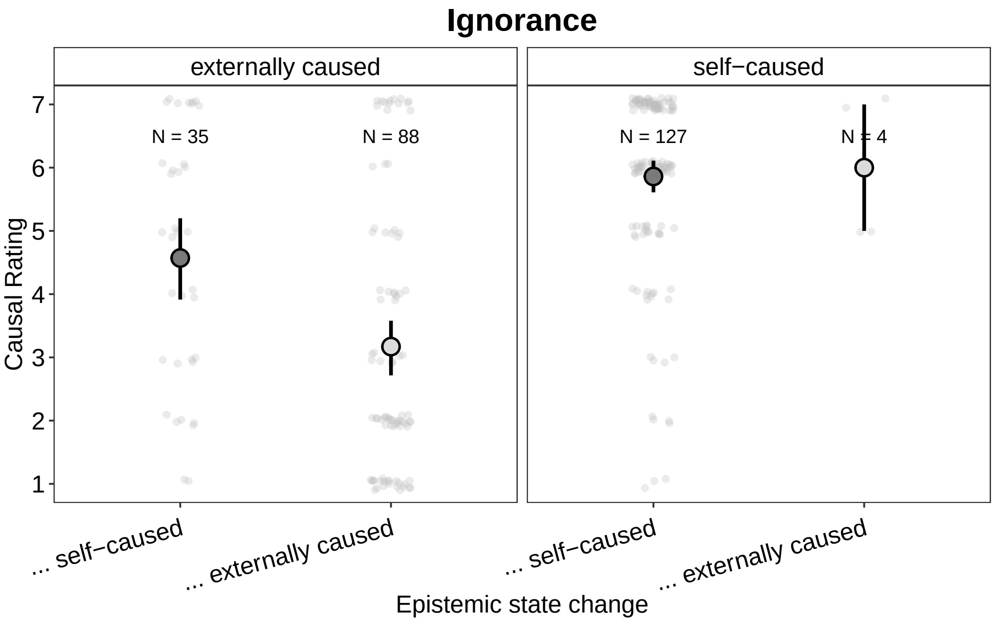


Figure 3.7: Experiment 2: Causal Ratings by Counterfactual Response Category. Causal Ratings of participants who gave imagined the agents' epistemic state change to be self or externally caused, split by ignorance condition

a self vs. externally caused epistemic change – independent of the experimental condition.

A subgroup analysis showed that the type of counterfactual response category participants chose predicted their causal ratings in addition to the ignorance condition, $\chi^2(1) = 11.43$; $p < .001$ [*between-subject contrast* $F(1) = 6.22$; $p = .01$]. Those people in the "externally caused ignorance" condition who still imagined a self-caused epistemic change gave a higher causal rating ($M = 4.57$, $SD = 1.97$, 95% CI [3.91, 5.23]) than those who imagined an externally caused epistemic change ($M = 3.17$, $SD = 2.10$, 95% CI [2.73, 3.61]), $t(234) = 3.65$, $p < .001$ [*between-subject contrast*: $t(121) = 2.63$, $p = .01$], (see Figure 3.8).

In the "self-caused ignorance" condition, there is no difference in ratings between participants who stated a external ($M = 6.00$, $SD = 1.15$, 95% CI [4.87, 7.13]) vs self-caused epistemic change (M =5.86, $SD = 1.46$, 95% CI [5.61, 6.11]), $t(250) = - 0.03$, $p = .98$, although it is important to note that only 4 people indicated the former type of response [*between-subject contrast*: $t(121) = -0.17$, $p = .86$].

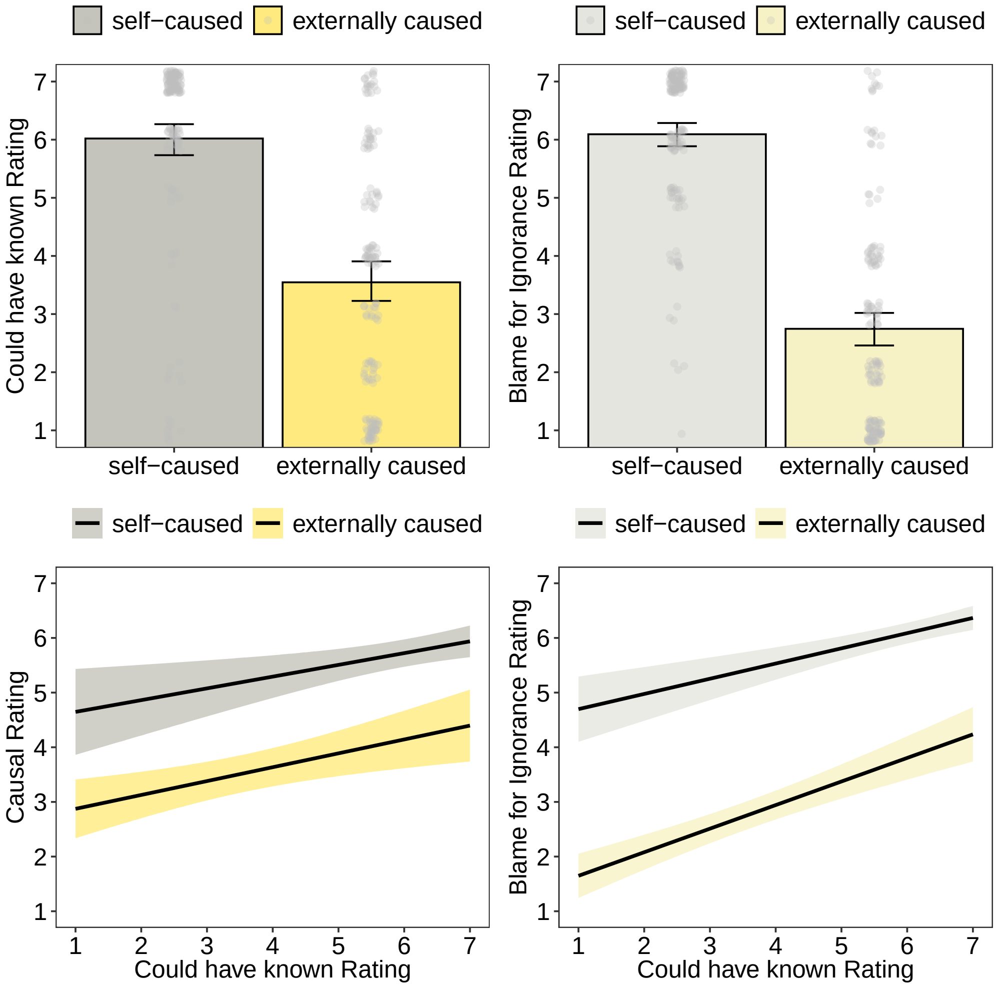


Figure 3.8: Experiment 2: Knowledge and Blame Ratings. Bar graphs depict means of i) "Could have known" ratings and ii) "Blameworthiness for Ignorance" ratings. Regression plots depict "Could have known" ratings as predictor for i) Causal ratings and ii) Blame for Ignorance ratings

3.9.3.4 Knowledge Rating

The condition under which ignorance came about significantly predicted people's modal judgement about the agent's epistemic state, $\chi^2(1) = 114.50$; $p < .001$ [*between-subject contrast*: $F(1) = 56.18$; $p < .001$]. People's agreed less that the agent could have known ($b = 2.46$, $SE = .19$, $t = 12.67$) when the agent was ignorant because of a technical default ($M = 3.55$, $SD = 2.11$, 95% CI [3.20, 3.90]) compared to ignorance caused by the agent themselves ($M = 6.02$, $SD = 1.69$ 95%, CI [6.72, 5.77]) (see Figure 3.7).

3.9.3.5 Blame Rating

Type of ignorance also influences people's judgement about the agent's blameworthiness for their ignorance, $\chi^2(1)$ = 237.15; $p < .001$ [*between-subject contrast*: $F(1) = 98.77$; $p < .001$], with people assigning less blame for the agent's ignorance when the ignorance was externally caused (*M* = 2.75, *SD* = 1.84, 95% CI [2.40, 3.10]) vs. self-caused (*M* = 6.09, *SD* = 1.27, 95% CI [5.84, 6.35]), (*b* = -3.34, *SE* = .16, *t* = 20.67). There was also a significant effect for scenario, $\chi^2(1)$ = 8.28; *p* = .02. Post-hoc t-tests revealed that blame for ignorance ratings in the hospital scenario are slightly higher compared to the garden scenario (*b* = 0.69, *SE* = .25, *t* = -2.76, *p* = .02).

3.9.3.6 Knowledge as Predictor

In a regression model that already includes the ignorance condition as a predictor, adding knowledge ratings improves the fit of the model for causal ratings $\chi^2(1)$ = 12.66; $p < .001$ (*b* = 0.20, *SE* = .06, *t* = 3.60) [*between-subject contrast*: $F(1)$ = 5.31; *p* = .02] as well as for blame ratings $\chi^2(1)$ = 59.00; $p < .001$ (*b* = 0.35, *SE* = .04, *t* = 8.07) [*between-subject contrast*: $F(1) = 37.58$; $p < .001$].

3.9.4 Discussion

The results of Experiment 2 show that the epistemic condition of ignorance influences people's causal judgements about the ignorant agent, their judgements about the mutability of the agent's epistemic state, as well as how blameworthy the agent is considered for their ignorance. Likewise, the epistemic condition also influences the target of intervention in people's counterfactual reasoning. Depending on whether the access to relevant information is prevented by an external cause or the agent's own actions, people differ in how likely they are to imagine an epistemic state that is brought about by the agent's action. Notably, a substantial proportion of people (25%) still indicated a self-caused epistemic change in the "externally caused ignorance" condition, mostly by referring to alternative information-seeking actions the agents could have done. When grouping participants by their counterfactual response, i.e. *how* the agent could have acquired knowledge, we found that participants who imagined an epistemic state change caused by the agent themselves gave higher causal ratings than those who imagined an externally caused knowledge acquisition. In cases where people's mental representation of the causal scenario included an alternative possible epistemic action for the agent, people also gave higher causal ratings. Thus, people's individual representation of a causal scenario influenced their causal ratings, mediated by the things they thought could have gone differently with respect to the agent's knowledge state.

In the next experiment, we wanted to follow up on the finding that an agent's epistemic action plays such a crucial role for people's causal judgement about their acting in ignorance. In particular, we were interested in whether it matters not only if the agent could have performed an epistemic action, but also how many epistemic actions it would have required them to obtain knowledge. According to most counterfactual theories, causal strength is generally sensitive to the number of changes that are necessary in order to render an outcome counterfactually dependent on a cause (Chockler & Halpern, 2004). Experiment 3 aimed to apply this idea of the number of epistemic actions that are required for epistemic change.

3.10 Experiment 3

In the third experiment, we aimed to assess judgments about agents who have access to knowledge, but vary in the number of actions they have to perform in order to obtain knowledge. Specifically, we wanted to test a case in which the agent only needs to perform few epistemic actions (Figure 3.9 A) or many epistemic actions (Figure 3.9 B) in order to acquire knowledge.

3.10.1 Participants and Design

We recruited 180 participants on Amazon Turk. 72 participants were excluded for not answering all eight comprehension check questions correctly, and one participant was excluded for providing a nonsensical counterfactual responses. The final sample consisted of 107 participants (M_{age} = 35.12, SD_{age} = 11.27, N_{female} = 27). We adopted a 2 ignorance (few actions vs. many actions) × 3 scenario ("hospital" vs. "garden" vs "bakery") design. 'Ignorance' was manipulated within participants

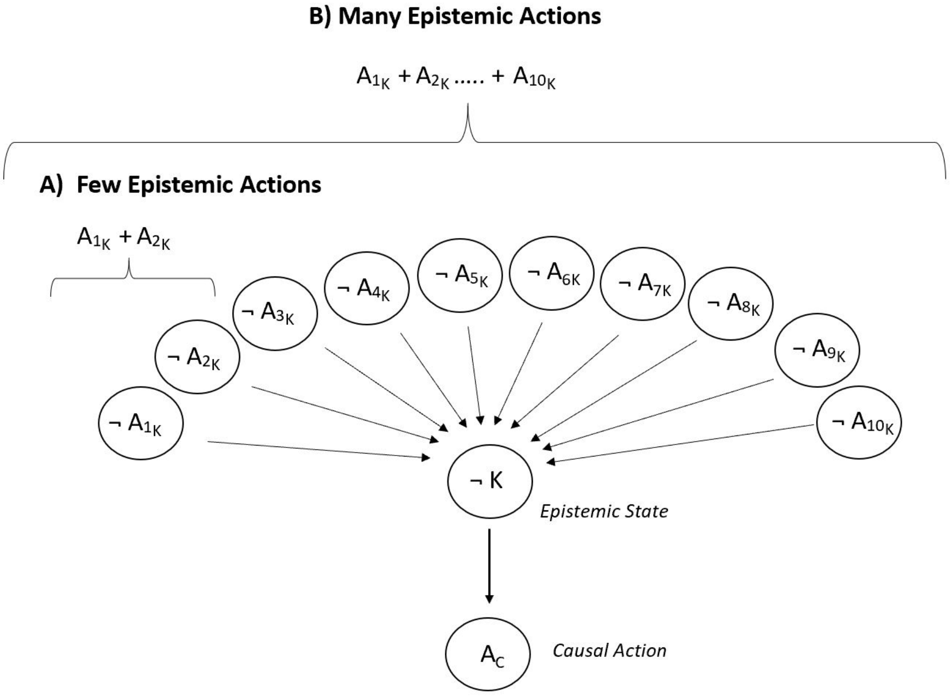


Figure 3.9: Experimental Conditions of Experiment 3. In the “Few Epistemic Actions” condition (A), it takes the agent two separate actions in order access information that would update their knowledge. In the “Many Epistemic Actions” condition (B), the agent needs to perform ten actions of inquiry in order to access the relevant information that would result in knowledge.

and ‘scenario’ was manipulated between participants.

3.10.2 Material

The main story followed Experiment 1 and 2: an agent unknowingly causes a harmful outcome by using a certain item. The information about the harmful properties of the item can be obtained by reading an e-mail with information about the item. However, this time it takes the agent to read through a few vs. many e-mails in order to obtain the relevant information.

As an example, in the “hospital” scenario, it is yet unknown that the drug “Afibo” causes leg cramps, but the hospital has ordered a series of drug tests that test for potential side effects.

> “[...] the fact that “Afibo” causes mild leg cramps in patients with blood type ‘AB-positive’ is still unknown. A standard procedure for hospitals

Table 3.1: Condition "Many actions"

Test Results									
E-mail 1	E-mail 2	E-mail 3	E-mail 4	E-mail 5	E-mail 6	E-mail 7	E-mail 8	E-mail 9	E-mail 10
no side effects	no side effects	no side effects	no side effects	no side effects	no side effects	no side effects	no side effects	no side effects	**effect**

Table 3.2: Condition "Few actions"

Test Results									
E-mail 1	E-mail 2	E-mail 3	E-mail 4	E-mail 5	E-mail 6	E-mail 7	E-mail 8	E-mail 9	E-mail 10
no side effects	**effect**	**effect**	**effect**	**effect**	**effect**	**effect**	**effect**	**effect**	**effect**

> is to let a specialised lab carry out a few tests on a new drug. Only a specialised lab can carry out the test that tests for a drug's side effect with certain blood types."

However, the sensitivity of the test for the side effect varies.

> *Few actions / Many actions* "This test for side effects with certain blood types is very sensitive [insensitive]. Statistically, 9 out of 10 tests detect "Afibo"'s side effect [only one out of 10 tests detects "Afibo"'s side effect] with blood type 'AB-positive'. How sensitive the test is is known to all doctors."

Ten tests are carried on the new drug, and depending on the sensitivity of the test, nine tests (few actions condition) or only one of the ten tests (many actions condition) do in fact detect the existent side effect. The results of each of the ten tests are conveyed to the doctors in ten separate e-mails (see Table 3.1 and Table 3.2).

> *Few actions / Many actions* "Ten of these tests have been carried out on "Afibo". One test result is negative [nine test results are negative], but

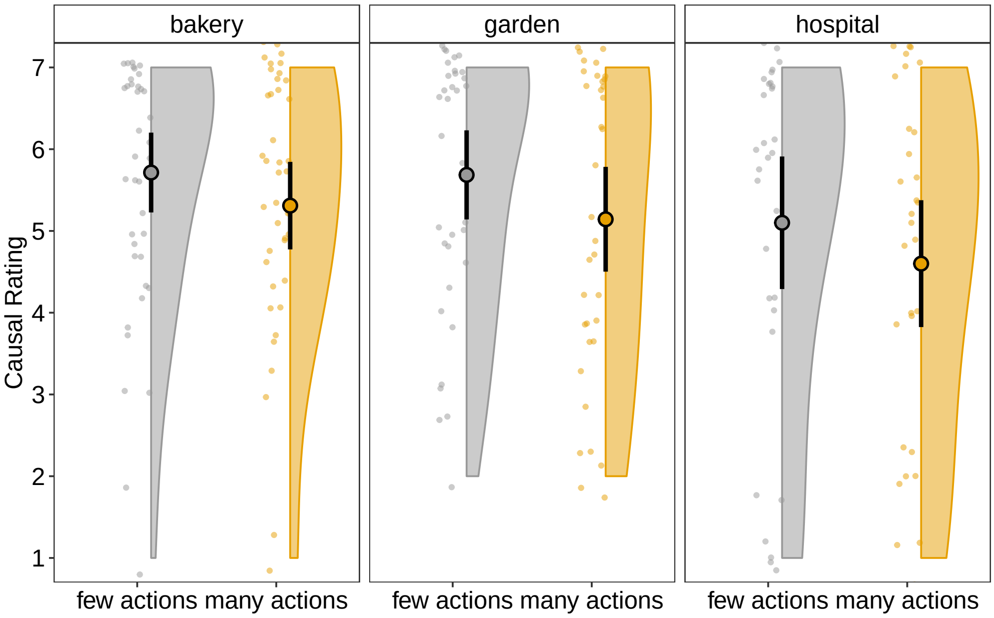


Figure 3.10: Experiment 3: Causal Ratings. Big dots are group means. Error bars depict 95% Confidence Intervals. Coloured backgrounds represent the probability distribution of the data, grey dots are individual participants' judgments jittered for visibility.

> nine tests finds evidence for the side effect [one test finds evidence for the side effect]. Each test result is sent to all doctors of the hospital in a separate e-mail."

In both experimental conditions, the doctor only reads the first in the row of ten e-mails in her inbox. This e-mail includes a negative test result (no effect detected) and therefore the doctor does not learn about the side effect. In consequence, in order to obtain the relevant information, it would have taken the doctor to read nine more e-mails in the "many actions condition" (only e-mail No '10' contains a positive test result, see Table 3.1), and at least one more e-mail in the "few actions" condition (e-mails No '2'-'10' contain a positive test result, Table 3.2). As before, the doctor prescribes the drug to a patient with the specific blood type and the patient is harmed.

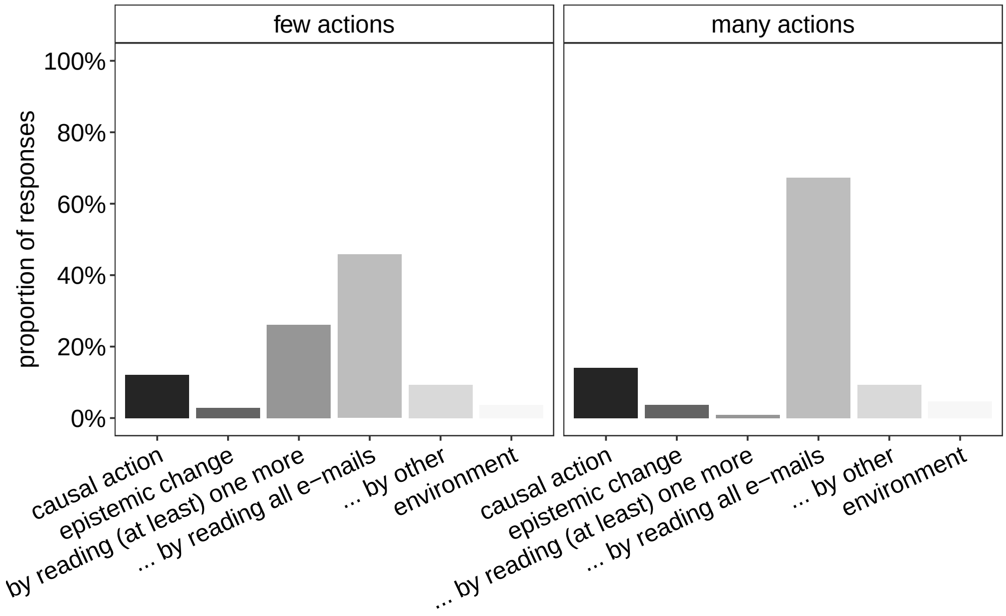


Figure 3.11: Experiment 3: Counterfactual responses. Proportion of responses in the four counterfactual response categories (“action“, “epistemic state change”, “epistemic state change by reading (at least) one more e-mail”, “epistemic state change by reading all e-mails”, “‘epistemic change by other”, “environment”)

3.10.3 Results

3.10.3.1 Causal Rating

The number of epistemic actions was a significant predictor for participants’ causal responses, $\chi^2(1) = 16.33$; $p < .001$, but only in the within-subjects condition [*between-subject contrast*: $F(1) = 0.03$; $p = .87$]. People saw the agent as less of a cause (b = -.48, *SE* = .12, t = -4.14) when many actions were necessary to obtain the relevant information and gain knowledge (*M* = 5.06, *SD* = 1.89, 95% CI [4.70, 5.20]) rather than few actions (*M* =5.53, *SD* = 1.77, 95%, CI [5.20, 5.81]) (see Figure 3.10). There was no significant effect of scenario (p = .23) [*between-subject contrast*: $F(2) = 1.24$; $p = .29$] and no interaction between ignorance and scenario (p = .87).

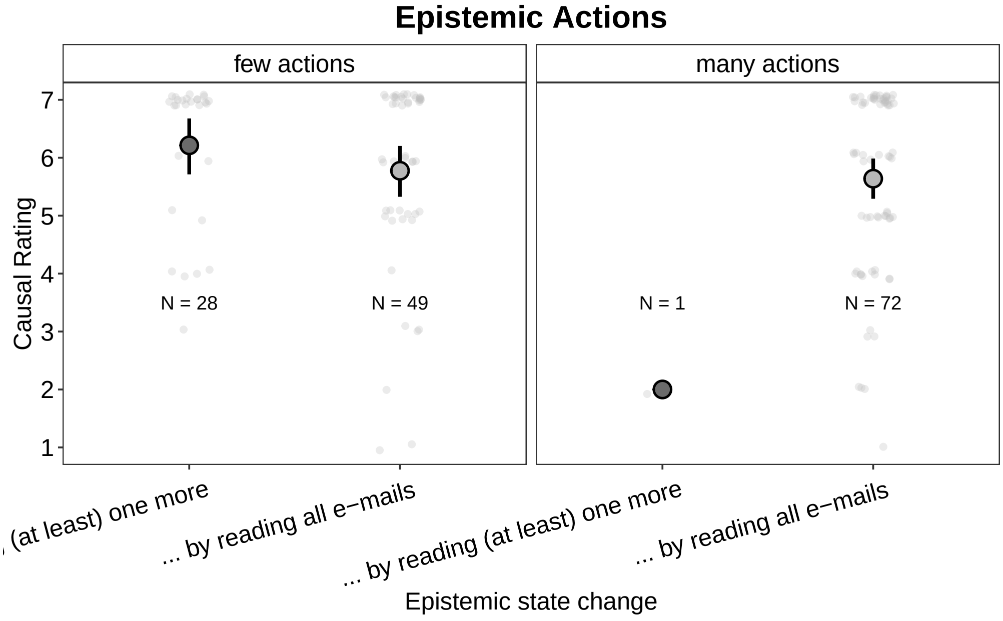


Figure 3.12: Experiment 3: Causal Ratings by Counterfactual Response Category. Causal Ratings of participants who gave imagined the agents' epistemic state change by few vs. many epistemic actions, split by ignorance condition

3.10.3.2 Counterfactual Responses

Based on participants' free text responses, we devised six coding rubrics for clustering the kinds of counterfactual changes participants imagined. Inter-rater agreement was at 90%.

- **"Action"**. Responses that described a change in the agent's action that caused the outcome, either by undoing the actual action (N = 28) (*"If Alex does not [sic] use the fertilizer splendor the bourbon roses would not have died"*).
- **"Epistemic state change (unspecified)"** Responses that suggested a direct change of the agent's knowledge about the item without specifying how (N = 7) (*"If Sandra knew there were nuts..').*
- **"... by reading (at least) one e-mail"** (N = 29), an epistemic state change caused by reading at least one more e-mail (*"If Alex had read more than 1 e-mail", "If Anne would have only looked at least at the next email..."*).

- **"... by reading all e-mails"** (N = 20), an epistemic state change caused by reading all available e-mails (*"If Dr. Smith had read the other 9 emails"*).
- **"... by other"** (N = 121), responses that suggested an epistemic state change that was not brought about by reading e-mails (*"If they were told that it contained traces of nuts...", "Sandra had been better instructed to read the e-mails thoroughly ..."*).
- **"Environment"** (N = 9), responses referring to changes in the environment that do not directly affect the agent's causal action, epistemic state or epistemic actions. (*"... the patient didn't have pre-existing health problems which required the intervention of medicine."*)

A multinomial logistic regression was performed to model the relationship between ignorance and response type, with "causal action" as reference category. A model with number of epistemic actions condition as predictor provided a significant fit for people's counterfactual responses $\chi^2(5)$ = 36.30; $p < .001$ R^2= .06 [*between-subject contrast*: $\chi^2(5)$ = 29.17; $p < .001$ R^2= .11]. Changing the epistemic action condition from "many actions" to "few actions" significantly increases the log odds of a response indicating a "(at least) one more e-mail response" (1% vs. 26%), compared to causal action responses (b = 3.48, OR = 32.31, SE = 1.10, z = 3.20, $p < .01$). However, a change in the epistemic condition from many to few epistemic actions did not significantly reduce the likelihood to indicate a response suggesting reading all e-mails (b = -0.24, OR = -0.28, SE = .42, z = -0.57, p = .57) (67% vs. 46%).

3.10.3.3 Counterfactuals: Subgroup-Analysis

As in Experiment 2, we broke down participant's causal judgments based on the kind of counterfactual response they gave in both conditions. In particular, we wanted to see whether participants who differed in terms of the number of epistemic actions indicated in their counterfactual response ("at least one more" vs. "all e-mails") also give different causal judgments. However, adding a predictor "coun-

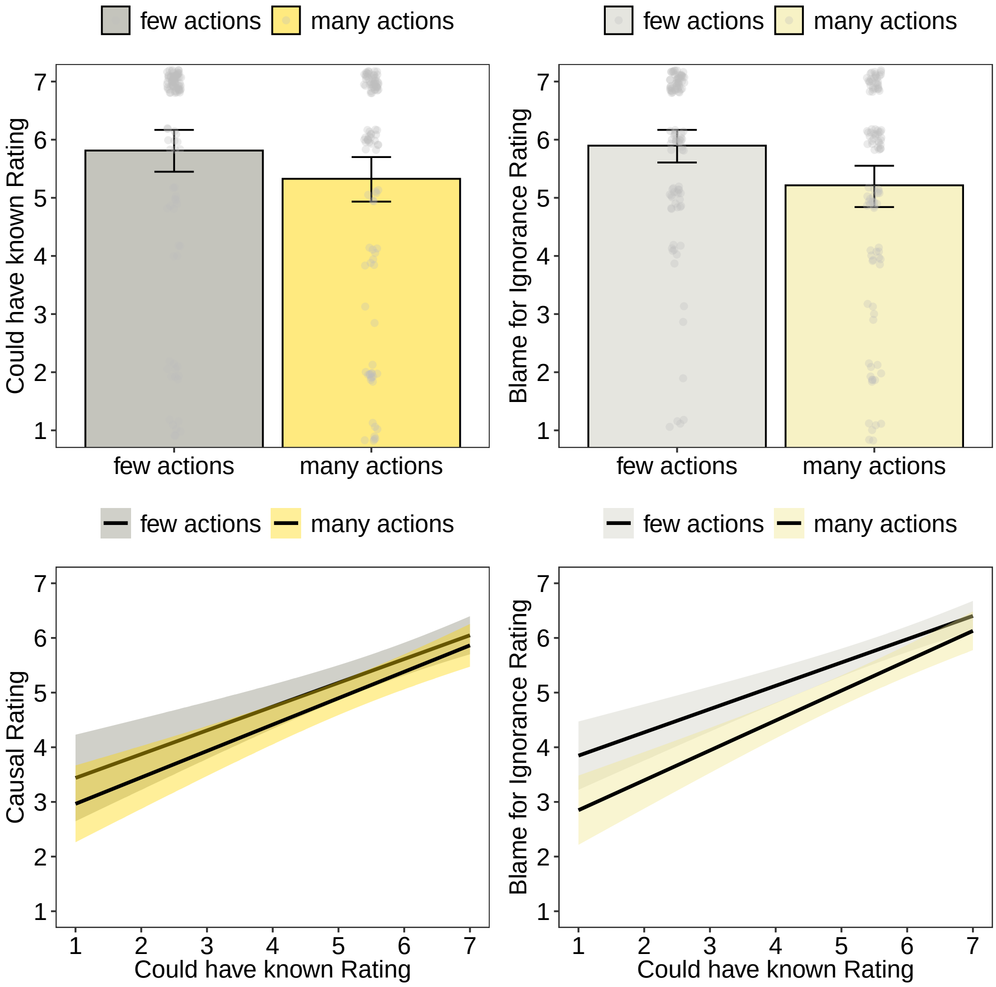


Figure 3.13: Experiment 3: Knowledge and Blame Ratings. Bar graphs depict means of i) “Could have known” ratings and ii) “Blameworthiness for Ignorance” ratings. Regression plots depict “Could have known” ratings as predictor for i) Causal ratings and ii) Blame for Ignorance ratings

terfactual response type” to a model already including ignorance condition did not provide a better fit for people’s causal judgments, $\chi^2(1) = .16$; $p = .69$, [*between-subject contrast*: $F(1) = 0.51$; $p = .48$] (see Figure 3.12).

3.10.3.4 Knowledge and Blame Ratings

Agreement ratings with the statement that the agent could have known about the harmful properties of their action were significantly influenced by the number of actions necessary for knowledge, $\chi^2(1) = 11.12$; $p < .001$. Ratings were lower when the agents would have needed to undertake more (M= 5.33, SD = 2.09, 95%

CI [4.97, 5.21]) compared few actions (*M* = 5.81, *SD* = 1.96, 95%, CI [5.48, 6.15]) (*b* = .49, *SE* = .15, *t* = 3.38) [*between-subject contrast*: $F(1) = 0.52$; $p = .47$]. Blame ratings were also influenced by the number of epistemic actions factor $\chi^2(1)$ = 29.11; $p < .001$. People assigned less blame in the "many actions" condition *M*= 5.33, SD = 2.09, 95% CI [4.97, 5.21]) compared the "few actions" condition (*M* =5.81, *SD* = 1.96, 95%, CI [5.48, 6.15]) (*b* = .68, *SE* = .12, *t* = 5.70) [*between-subject contrast*: $F(1) = 0.55; p = .46$]. In addition to the epistemic action condition, knowledge rating was a significant predictor for people's causal judgements $\chi^2(1)$ = 52.89; $p < .001$ (*b* = 0.41, *SE* = .06, *t* = 6.74) [*between-subject contrast*: $F(1)$ = 32.70; $p < .001$] as well as for blame ratings $\chi^2(1) = 59.00$; $p < .001$ (*b* = 0.43, *SE* = .05, *t* = 1.22) (see Figure 3.13) [*between-subject contrast*: $F(1) = 53.37$; $p <$ .001].

3.10.4 Discussion

In Experiment 3, we found that the number of actions that an agent needs to perform in order to change their knowledge state influences how causal people judge them for the unknown outcome of their action. However, we could find this effect only as within-contrast. That is, only when considering the responses that each participant made to both the "few epistemic actions" and the "many epistemic actions" condition, we observed a difference in their judgements about causation, changeability of epistemic state, and blame for ignorance. In the scenarios of our experiment, the epistemic action is always of the same kind – reading an e-mail – , but varies in the number of actions required in order to lead to the information. It is possible that without a direct comparison contrast, i.e. without being able to compare the number of 2 vs. 9 required actions ("If the agent had read one more vs. all emails, ...") , people do not naturally represent these as scenarios as including few vs. many epistemic actions (Schaffer, 2005). Rather, people might in the first instance represent these as a general action "reading e-mails", and hence as a single epistemic action variable that they intervene on. This might explain why we only found the effect as a within-subject contrast. Stronger, more intuitive manipulations of different requirements of epistemic actions and epistemic effort might help to show these

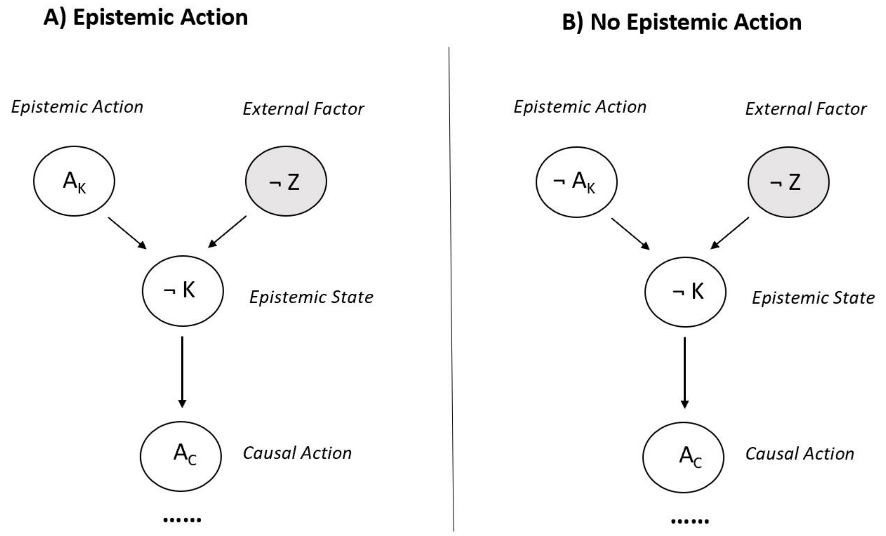


Figure 3.14: Experimental Conditions of Experiment 4. In the 'Epistemic Action' condition (A), the agent performs an epistemic action (reading e-mail), but the external conditions for information acquisition are not given (information not included in the e-mail) and they remain ignorant. In the 'No Epistemic Action' condition, the agent does not perform the epistemic action (not reading e-mail), the external condition for information acquisition is not given (information not included in the e-mail), and the agent continues to be ignorant.

differences as between-contrasts as well.

The final experiment in this chapter aims to investigate the consequence of epistemic actions under different circumstances. According to counterfactual theories of causation, causality is determined by the counterfactual dependence of the outcome on the candidate cause in the actual world, but also under different 'contingencies', e.g. when background circumstances are different (Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2014; Halpern, 2016). In Experiment 4, we want to apply this notion of counterfactual dependence under different contingencies to the epistemic state of a causal agent. That is, we wanted to test whether people take into account agents' epistemic actions, even if the agent's actions do *not* lead to the acquisition of knowledge in the actual scenario, but would have under different circumstances.

3.11 Experiment 4

In our last experiment, we aimed to test whether people take into account if an agent performs an epistemic action, i.e. whether they aim to acquire information, even if this epistemic action is without consequence for their knowledge about the outcome. In these scenarios, the external factor that is necessary for knowledge acquisition is not given – the crucial information about negative consequences is missing in an e-mail about the relevant item. We vary whether the agent performs an epistemic action (reading e-mail) which does *not* lead them to obtain the relevant information given that it is missing (see Figure 3.14 A), or whether they do not even perform the epistemic action (see Figure 3.14 B).

3.11.1 Participants and Design

We recruited 171 participants on Amazon Mechanical Turk. 34 participants were excluded for failing one or more of the four comprehension check questions, and 2 participants were excluded for providing a non-sensical counterfactual response, leaving a final sample size of $N = 133$ (M_{age} = 38.36, SD_{age} = 11.38, N_{female} = 57, 1 = unidentified). We adopted a 2 ignorance (information search vs. no information search) $\times$ 3 scenario (“hospital” vs. “garden” vs. “bakery”) design. ‘Information acquisition’ was manipulated within participants and ‘scenario’ was manipulated between participants.

3.11.2 Material

In the frame story of Experiment 4, an email about the relevant item is (successfully) sent to the agent. However, in this e-mail, the crucial information about the harmful property of the item is missing:

> “[...] in this e-mail, the paragraph on side effects is missing. The e-mail does not contain the information that“Corus” causes mild leg cramps in patients with blood types ‘B-negative’.”

We then varied whether the agent read (“information-seeking”) or did not read the e-mail (“not information-seeking”). As before in Experiment in 2, in both conditions the agent unwittingly applies the harmful item with negative consequences.

3.11.2.1 Knowledge and Blame for Ignorance

These ratings and response measures were obtained as in Experiment 2 and 3.

3.11.2.2 Forward-looking causal judgments

In order to investigate whether people's causal judgments in the actual scenario is related to how they would judge about the agent if circumstances were different, we included a follow up scenario. Participants were prompted to imagine a future scenario in which there is a new pain killer "Innohep" ('bakery' scenario: flour brand, 'garden' scenario: weed killer) in hospitals. However, this pain killer causes nausea in patients who take beta-blockers. As usual, an e-mail has been sent out to all doctors, introducing the new pain killer. However, this time the e-mail *does* include the information that this pain killer causes nausea in patients taking beta-blockers. Participants were then asked to estimate the likelihood that the agent from the "information-seeking" condition and the agent from the "non-information seeking" condition would read that e-mail in this future scenario: "How likely is it that Dr Jones [Dr Smith] would check the e-mail of the pharmacy manager about 'Innohep' "? (0 - "Extremely unlikely"; 100 - "Extremely likely"). In addition, they were asked about the likelihood of a bad outcome given that either agent would be in charge of a patient with the sensitive condition: "How likely is it that a patient who takes beta-blockers would suffer from nausea if Dr Jones were treating this patient [Dr Smith were treating this patient]" (0 - "Extremely unlikely"; 100 - "Extremely likely"). These two follow up questions allowed us to test whether differences in causal judgments in the actual scenario might correspond to what would have happened in a different epistemic context, e.g. if the e-mail would contained the relevant information.

3.11.3 Results

3.11.3.1 Causal Ratings

The "information seeking" factor, i.e. whether the agent read the e-mail or not, was a significant predictor for participants' causal responses, $\chi^2(1) = 38.91$; $p < .001$ [*between-subject contrast*: $F(1) = 12.00$; $p < .001$]. People judged the agent to be

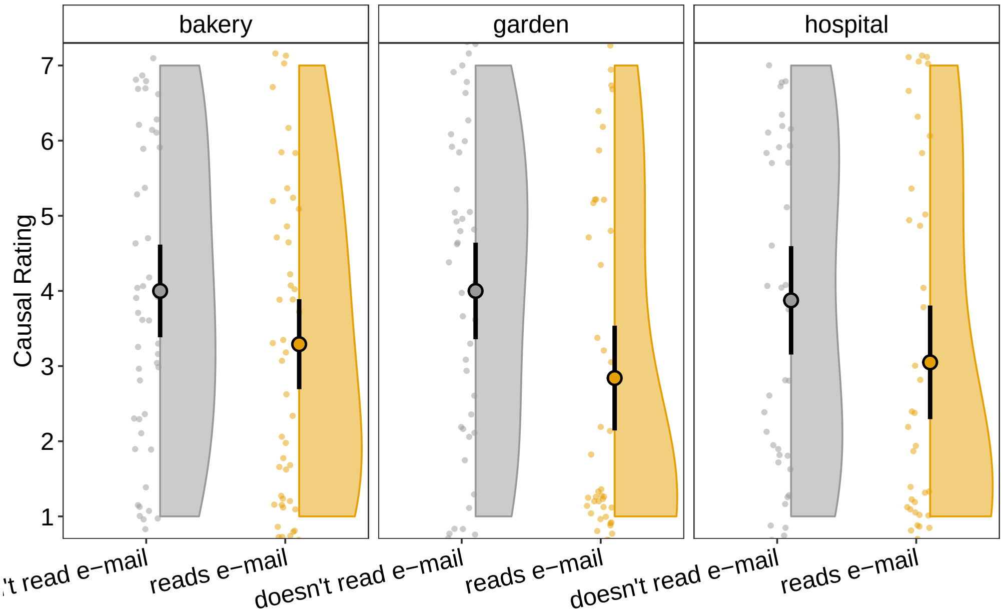


Figure 3.15: Experiment 4: Causal Ratings. Big dots are group means. Error bars depict 95% Confidence Intervals. Coloured backgrounds represent the probability distribution of the data, grey dots are individual participants' judgments jittered for visibility.

less of a cause (b = -.71, SE = .22, t = -3.25) when the agent read the e-mail with the missing information (M = 3.10, SD = 2.24, 95% CI [2.72, 3.48]) than if they did not (M = 3.96, SD = 2.14, 95%, CI [3.60, 4.33]) (see Figure 3.15).

3.11.3.2 Counterfactual Responses

Clustering participant's responses revealed the following categories:

- **"Causal Action"** Responses that described a change in the agent's action that caused the outcome, either by undoing the actual action (N = 14)
- **"Epistemic state change (unspecified)"** Responses that suggested a direct change of the agent's knowledge about the item without specifying how (N = 17).
- **"... by information"** (N = 90), through the availability of the relevant information in the e-mail (' '*If the email had contained the warning...*") (N = 90),

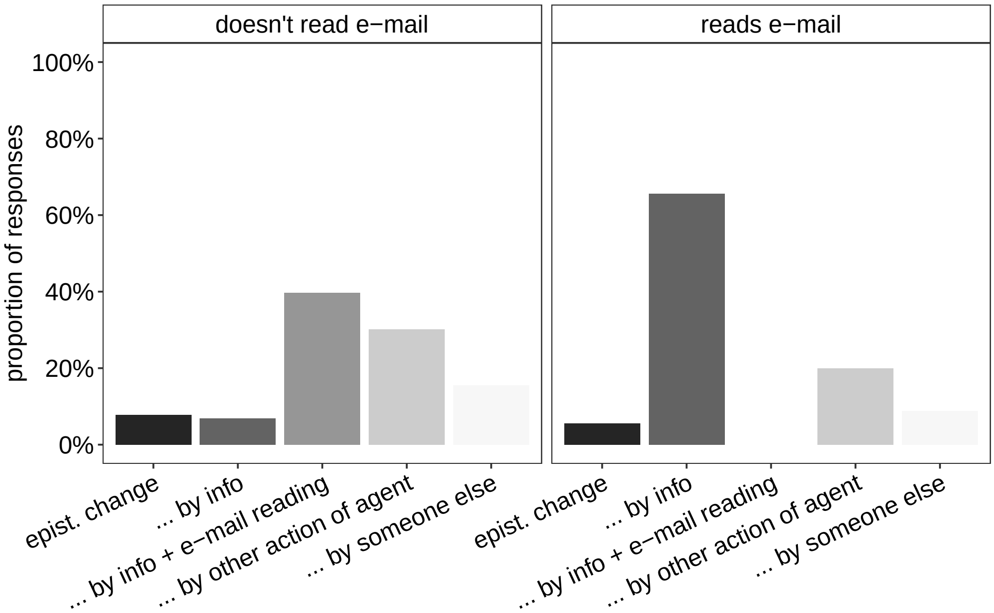


Figure 3.16: Experiment 4: Counterfactual responses. Proportion of responses in the five counterfactual response categories (“epistemic state change, unspecified“, “epistemic state change by addition of information in e-mail”, “...by addition of information and agent reading e-mail”, “...by another epistemic action of the agent” and “.... by another agent”)

- **“... by reading the e-mail” (*N* = 4)**, the agent reading the e-mail (*“If Anne would have read the e-mail”*).
- **“... by info + reading the e-mail”** (*N* = 46), the e-mail containing the relevant information and the agent reading the e-mail (*“If the email contained a warning about the effects of the new fertilizer AND Bob read the email...”*)
- **“... by other action of focal agent ”** (*N* = 60), by some other, alternative action of the agent that would have led them to acquire knowledge (*“If Sandra had done more research...”*)
- **“... by someone else”** (*N* = 29), by a third party-agent informing the focal agent about the harmful effect (*”If the company had told Sandra that the product had traces of walnuts...”*).
- **“Environment”** (*N* = 7) responses referring to changes in the environment.

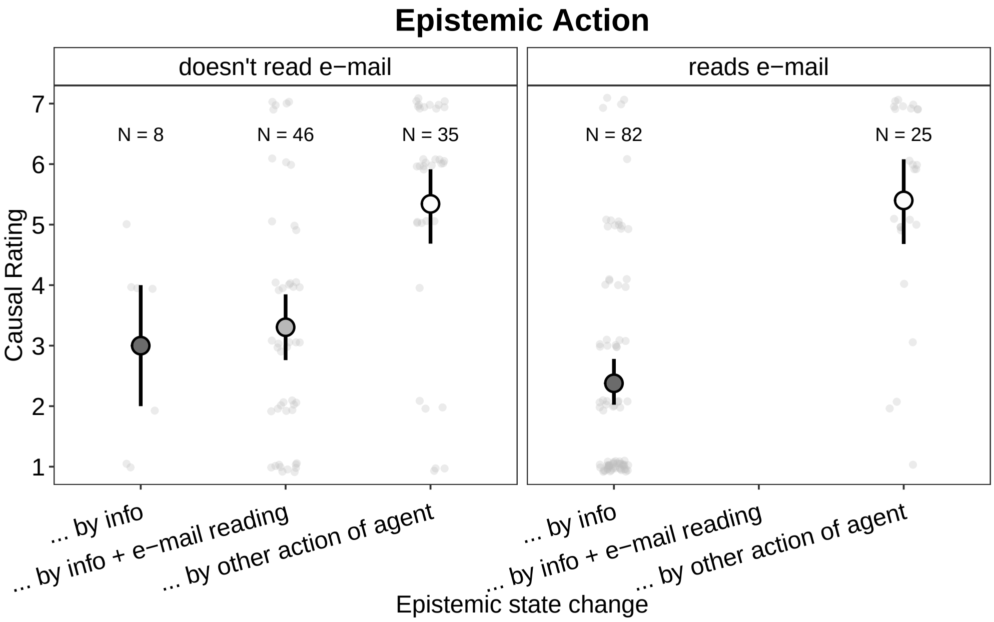


Figure 3.17: Experiment 4: Causal Ratings by Counterfactual Response Category. Causal Ratings of participants who gave imagined the agents' epistemic state change to be caused by i) the e-mail containing the relevant information, ii) the e-mail containing the relevant information and the agent reading it and iii) by some alternative epistemic action by the agent, split by ignorance condition

Inter-rater agreement of clustering participants' responses was at 90%. In order to keep the analysis of counterfactual responses concise, we excluded those response categories that had less than 5% of participants' responses across both "information-seeking" conditions: "causal action", "...reading the e-mail" and "environment".

The "unspecified epistemic state change" response category was chosen as a reference category. The information acquisition condition significantly predicted people's counterfactual responses $\chi^2(4)$ = 137.84; p < .001, R^2= .20 [*between-subject contrast*: $\chi^2(4)$ = 65.49; p < .001, R^2= .20]. When the agent did not read the e-mail, people were less likely to indicate a change that consisted in the addition of *just* the missing information in the e-mail ("...by info") (b = -2.58, OR = .08, SE = .63, z = -4.12, p < .001) (9% vs. 66%), compared to a change in just the epistemic state (see Figure 3.16) (b = -1.80, OR = -.16, SE = .91, z = -1.96, p < .001).

3.11.3.3 Counterfactuals: Sub-group Analysis

As in the studies before, we analysed people's causal judgments in dependence of which kind of counterfactual response they gave. In particular, we were interested in comparing those participants who imagined an epistemic state change caused by an alternative action of the agent ("... by other action of agent") to those participants whose responses corresponded to the manipulations in the experiment ("by info", "by info + e-mail reading"). Adding "counterfactual response type" as a predictor to a model including the "epistemic action" factor significantly improved the fit of the model for causal judgments, $\chi^2(1)$ = 41.48; p < .001 [*between-subject contrast*: $F(1)$ = 15.34; p < .001]. In the "doesn't read e-mail" condition, people who imagined the agent to perform an alternative action in order acquire knowledge gave higher causal ratings (*M* = 5.34, *SD* = 1.94, 95% CI [4.70, 5.99]) than those who indicated that the e-mail could have had the relevant information and the agent could have read it, (*M* =3.30, *SD* = 1.95, 95% CI [2.74, 3.87]), *t*(166) = -4.22, p < .001 [*between-subject contrast*: *t*(87) = -3.04, p < .01] (see Figure 3.17). The difference in causal ratings between "by info" (*M* =2.38, *SD* = 1.80, 95% CI [2.00, 2.77]) and "by other action of agent" (*M* =5.40, *SD* = 1.78, 95% CI [4.70, 6.10]) responders was also significant in the "reads e-mail" condition, *t*(175) = -6.76, p < .001 [*between-subject contrast*: *t*(87) = -4.84, p < .001].

3.11.3.4 Knowledge and Blame Ratings

Information-seeking behaviour significantly predicted modal judgments about the agent's epistemic state $\chi^2(1)$ = 14.08; p < .001 [*between-subject contrast*: $F(1)$ = 4.38; p = .03], as well as blameworthiness for ignorance $\chi^2(1)$ = 54.47; p < .001 [*between-subject contrast*: $F(1)$ = 18.42; p < .001] (see Figure 3.18). The agent who did not read the e-mail containing missing information was judged to could have known about the relevant information to a greater extent (*M* = 3.43, *SD* = 2.24, 95% CI [3.07, 3.80]) and to blame more for their ignorance (*M* = 3.49, *SD* = 2.13, 95% CI [3.13, 3.85]) than the information-seeking agent ("Could have known": *M* = 2.93, *SD* = 2.24, 95% CI [2.55, 3.31]; "Blame": *M* = 2.47, *SD* = 2.07, 95% CI [2.09, 2.85])

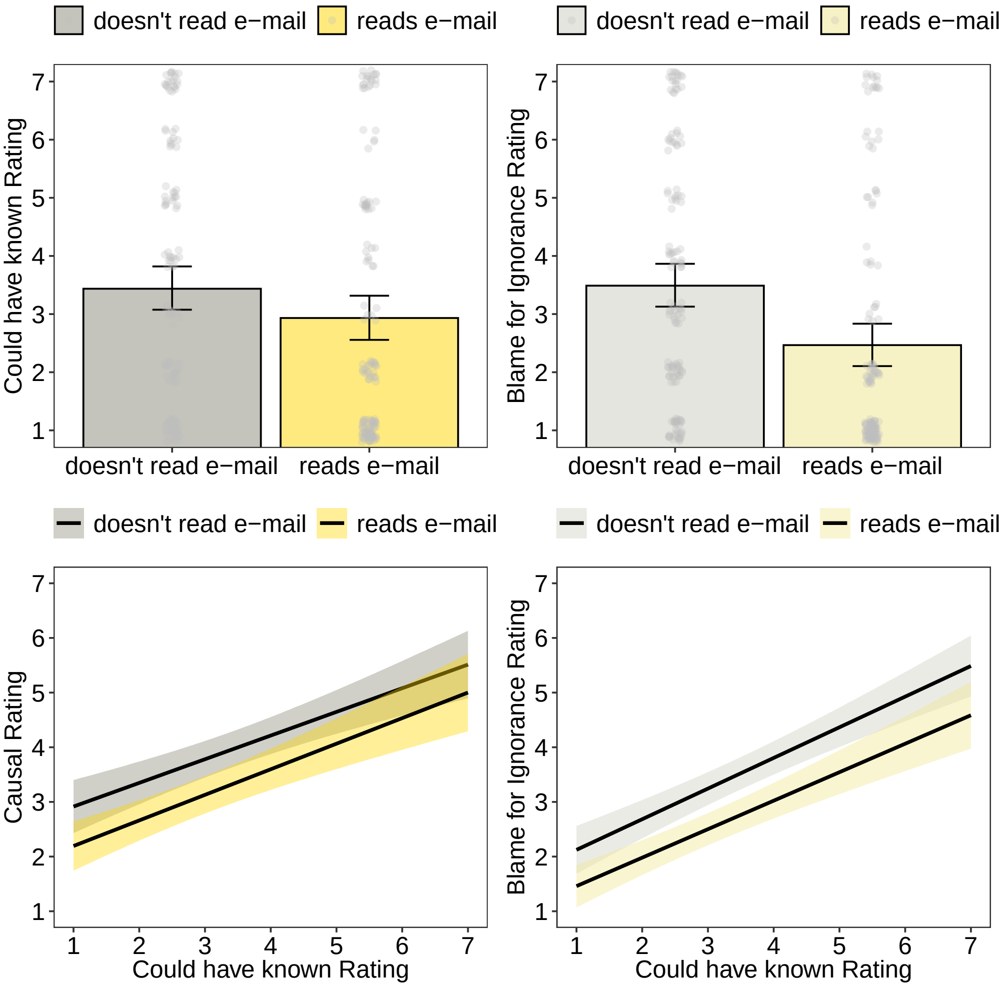


Figure 3.18: Experiment 4: Knowledge and Blame Ratings. Bar graphs depict means of i) "Could have known" ratings and ii) "Blameworthiness for Ignorance" ratings. Regression plots depict "Could have known" ratings as predictor for i) Causal ratings and ii) Blame for Ignorance ratings.

Adding knowledge rating as a predictor significantly improved a model that already contained the "epistemic action" condition for people's causal ratings, $\chi^2(1) = 63.27$; $p < .001$ ($b = 0.44$, $SE = .06$, $t = 7.17$) [*between-subject contrast*: $F(1) = 31.27$, $p < .001$] as well as blame ratings,$\chi^2(1) = 105.64$; $p < .001$ ($b = 0.55$, $SE = .05$, $t = 10.30$) [*between-subject contrast*: $F(1) = 72.96$, $p < .001$].

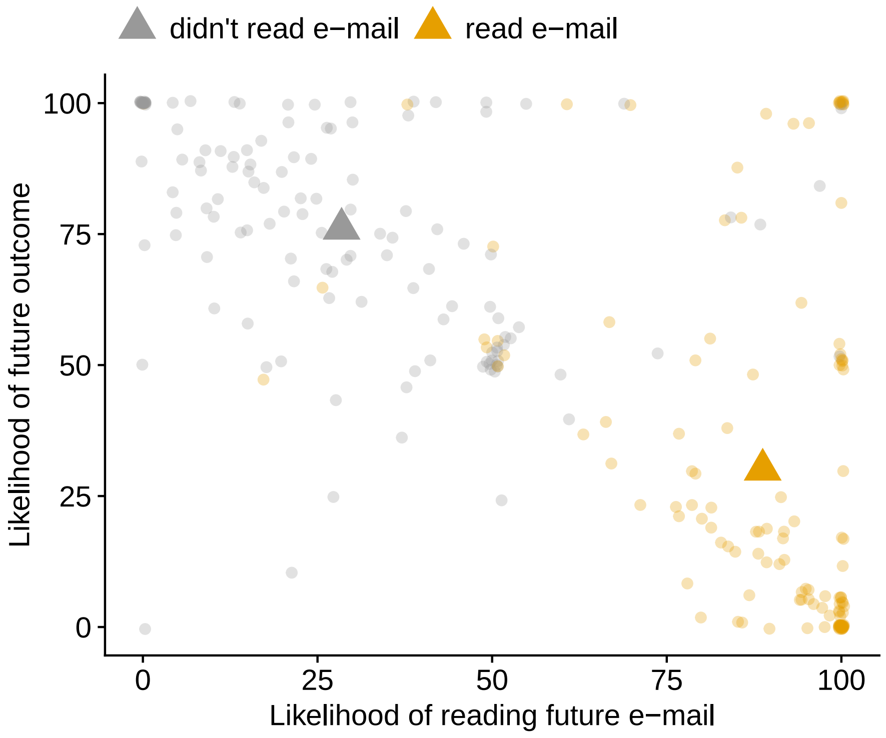


Figure 3.19: Experiment 4: Forward looking Causal Judgments. Likelihood ratings of the agent reading an e-mail about a future item in future (X-Axis) and Likelihood ratings of a future outcome (Y-Axis). Dots are individual participants' judgments, triangles are group means, grouped by previous epistemic action (reading vs. not reading e-mail).

3.11.3.5 Forward-looking Causation: Rating of Likelihood of Epistemic Action and Outcome

The epistemic action condition, i.e. whether in the the agent read the e-mail that was missing crucial information in the first experimental scenario, was a significant predictor of how likely participants rated the agent to read the e-mail in a future scenario $F(1) = 511.67$; $p < .001$ (see Figure 3.19). When the agent had read the e-mail before, they were seen as more likely to do the same in a future situation (M = 88.57, SD = 17.93, 95% CI [85.52, 91.62]) than if they had not read the e-mail before (M = 28.98, SD = 24.55, 95% CI [24.81, 33.15]). The previous epistemic action also affected how likely people judged the bad outcome to happen in the future scenario, $\chi^2(1) = 163.51$; $p < .001$. Correspondingly, people judged the outcome as less likely to happen when the agent had read the e-mail in the scenario before (*M* = 30.46, *SD* = 34.62, 95% CI [24.57, 36.34]) compared to when the agent had not attempted to acquire knowledge before (*M* = 75.84, *SD* = 21.78, 95% CI [72.14, 79.54]).

3.11.4 Discussion

In our final experiment, we found evidence that people take into account an agent's epistemic actions 'under different contingencies'. Experiment 4 showed that an agent who unsuccessfully attempts to acquire knowledge because of a lack of relevant information is still seen as less causal for the unforeseen outcome than an agent who does not attempt to do so, even if the attempt would be have been equally unsuccessful. We also found this difference in people's judgments about blame as well as their judgments about whether the agent *could* have known about the outcome, both as within- and between-contrast. The fact that information-seeking is taken into account for the perceived causal strength of the agent likely results from people integrating alternative scenarios with different circumstances into their counterfactual thinking. In a world in which the e-mail had contained the relevant information, the agent who read the e-mail would have found out about the negative outcome, and the outcome would potentially not have occurred. People's forward-looking causal judgments in the follow-up scenario supported this. Based on the agents' prior epistemic (non)-actions, people predicted the agent who had read the e-mail before to do so again, and in consequence judged the likelihood of a similar, future outcome to be lower. The likelihood of the outcome here includes some natural uncertainty about whether the epistemic state change also results in the absence of the causal action. It is however significantly predicted by previous epistemic actions. Hence, people do not only assess the dependence of an outcome on an action against a variety of background circumstances, but also the dependence of outcome on epistemic actions under different epistemic contexts.

3.12 General Discussion

Did Romeo cause Juliet's death? According to research in causal cognition, Romeo's ignorance of the consequence of his actions reduces his perceived causal contribution to Juliet's death. The kind of knowledge that an agent possesses, and moreover, the kind of knowledge that an agent does *not* possess, influences how causal the agent is perceived for an outcome. In this chapter, we were interested in the role that agent epistemic states play in causal judgments. In particular, we pursued the hypothesis that the influence of ignorance on causal judgements reveals something about the way people think about agent causality. Specifically, the proposal we make in this chapter aims to explain the influence of epistemic states on causal judgements by reference to counterfactual reasoning (Gerstenberg et al., 2020; Halpern, 2016; Halpern & Hitchcock, 2015). We argue that people represent epistemic variables in their causal representations of the world, and use these as points of intervention in counterfactual reasoning. More precisely, people imagine a change in the agent's epistemic state or epistemic actions when thinking about how things could have gone differently. In four experiments, we found that the extent to which people judged an ignorant agent to be causal corresponded to the point of epistemic interventions in their counterfactual reasoning. In addition, people's judgments about whether and how easily an agent could have known about the outcome of their action predicted their causal judgements, and also their judgments about blameworthiness for ignorance. Epistemic states and actions play a role for causal reasoning even when an agent could not have change their epistemic state in the actual world, but would have acquired knowledge in an alternative world with different epistemic conditions.

Drawing on causal model theory (Sloman & Lagnado, 2015), we made a first attempt at explaining how the role of epistemic states could be integrated into counterfactual frameworks. In principle, however, we think that our findings might be explained by a variety of accounts that draw on the notion of counterfactual reasoning. In the first part of our General Discussion, we will briefly sketch how such related explanations might look. Crucially, however, we will also revisit our findings with respect to theories of blame for ignorance, and the general role of blame in these cases. This will be discussed in the second part of the General Discussion. Lastly, we want to touch upon some potential implications that we think this work could have for the debate around normative vs. non-normative accounts of causation .

3.13 Integrating Epistemic States into Formal Frameworks of Causation

3.13.1 Halpern Hitchcock's Means Rea and Possible World Ordering

Halpern and Hitchcock (2015) extend their formal framework for incorporating defaults, typicality, and normality by suggesting an ordering of possible worlds that is based on the normality of variables. One such example is the role of different mental states, or "mens rea", for intervening causation in the law (Knobe & Shapiro, 2021). If Anne negligently spills gasoline and Bob carelessly throws a cigarette on the floor, Anne is legally determined as the cause for the resulting fire (Hart & Honoré, 1959/1985). However, her causal status is overriden by Bob's action if Bob throws the cigarette maliciously rather than negligently, with Bob now being determined as the cause. In line with the different types of "mens rea" (Kneer & Bourgeois-Gironde, 2017), the agents' mental states can be ordered according to their degree of culpability, i.e. *carelessness* < *negligence* < *maliciousness*. Drawing on work on the influence of normality on counterfactual reasoning (Hitchcock & Knobe, 2009), Halpern and Hitchcock (2015) propose that possible worlds are ordered by their status of normality, and that the the most "normal" (here: prescriptively normal) comparison contrast is prioritised. The pattern of intervening causation can be explained if in addition to the agents' actions, their mental states are represented as variables that can take the values 1 or 0 (e.g. BM = 1 – Bob is malicious, BM = 0 – he is not). The most "normal" contrast to the first scenario would be one in which Anne wasn't negligent and hence had not acted (AN = 0, but Bob is still careless, BC = 1), rendering her a cause. However, since maliciousness is a more culpable state than negligence, the prioritised contrasted possible world for the second scenario would be one in which Bob is not malicious (BM = 0, but

Anne is still negligent, AN = 1), making Bob the cause of the fire.

In Halpern and Hitchcock (2015)'s framework, an agent's action is automatically undone with the change of mental state, i.e. Anne not being negligent involves Anne not spilling the gas. They also sometimes refer to "Bob's malice" or "Anne's negligence" as the cause of the fire. In general, however, their proposal of representing mental states in structural equation models of causation is similar to the account we propose here, but differs in two aspects. We argue that people selectively intervene on epistemic states, and that there is more uncertainty about the outcome being different if the intervention targets a prior variable than the causal action. In addition, the account of Halpern and Hitchcock (2015) only takes into account mental states in their function of providing an normality ordering over possible worlds. Our account has the advantage of arguing for the general role of epistemic states, e.g. by pointing out suitable targets of intervention, independent of how normal or abnormal the epistemic state is (although these two features might sometimes align). We predict that the influence of epistemic states on causal judgments, such as reduced causal attributions to ignorant agents, is independent of how abnormal their lack of knowledge or epistemic actions is perceived.

3.13.2 Structural Model Account of Blame

Chockler and Halpern (2004) extend Halpern (2008)'s framework of causation for a definition of blame that takes into account an agent's epistemic state. According to this approach, blame is relative to an agent's epistemic state, which is taken to be a set of situations that the agent considers possible before the action is performed, together with a probability on them. An agent's blameworthiness for an action is hence their expected degree of responsibility for the outcome summed over all possible scenarios. To put this more concretely, in case of a doctor who is completely ignorant about the side effect of the drug they give to a patient, their degree of blame is 0 since they would not have expected to cause the effect. The measure of blame allows us to incorporate different types of uncertainty. Imagine that the doctor knows that if the patient has skin condition A, the drug will produce a side effect with a likelihood of 50%, but if he has skin condition B or C, the side effect will occur with a likelihood of only 10%. Each skin condition is equally likely to occur. According to Chockler and Halpern (2004), the doctor's degree of blame for the side effect here would be $\frac{1}{3} \times \frac{1}{2}$ (responsibility in condition A) $+ \frac{1}{3} \times \frac{1}{10}$ (responsibility in condition B) $+ \frac{1}{3} \times \frac{1}{10}$ (responsibility in condition C) = .23.

Crucially however, Chockler and Halpern (2004)'s measure of degree of blame considers the epistemic state of the agent *before* the action was performed. Chockler and Halpern (2004) argue that people will update their blame judgment about the doctor after the treatment, when her knowledge about the causal effects and structure has changed. While Chockler and Halpern (2004) only consider epistemic states for blame (more on the blame vs. cause distinction later), their theory departs from our proposal by considering the situation prior to the outcome. The account we propose in contrast considers the agent's causality *for* the outcome, relative to the agent's epistemic state at the time of the action. Chockler and Halpern (2004)'s account, however, raises the important point to integrating degrees of epistemic uncertainty rather than just binary epistemic states like knowledge vs. ignorance of a fact.

3.13.2.1 Normality and Sampling Propensity

Icard et al. (2017) draw on the process of counterfactual sampling in order to explain the role that normality plays in causal judgements. According to their account, causal strength is assessed by stochastically sampling counterfactuals and using these to determine the extent to which a factor is causally relevant to a given outcome. How likely people sample a certain counterfactual – the sampling propensity – is influenced by aspects of normality, both prescriptive as well as descriptive. Sampling propensity is directly proportional to the normality of a counterfactual (Icard et al., 2017; Kominsky et al., 2015; Phillips et al., 2019), with normal counterfactuals more likely to be sampled than abnormal ones. People are more likely to sample counterfactuals in which an abnormal factor is absent compared to counterfactuals in which the normal causal factor is absent, leading to an increased perceived causal strength of abnormal causes.

Our account does not draw on the specific role of normality, but can also be spelled out in terms of probabilistic sampling processes (Icard, 2016). People might be less likely to sample counterfactual scenarios in which the doctor does not prescribe the drug if the doctor is ignorant compared to when they know about the side effects. According to this account, similarly to normality, epistemic states influence the sampling propensity of counterfactual worlds in which the agent refrains from acting, based on probabilistic (and potentially normative) assumptions about whether an agent possessing (or lacking) relevant knowledge about the properties of their action performs this action.

3.13.3 Process-theory of Causality: Force Theory of Causation

The force dynamics model characterizes causation as a pattern of forces and a position vector (Talmy, 1988; Wolff, 2007). Often contrasted with counterfactual dependence theories, this theory distinguishes between different causal relations by specifying them in terms of configurations of forces. Force-theories of causation assume that agent causation is in some way modelled after physical causation (Talmy, 1988; Wolff, 2007). These theories suggest that mental states like intentions or desires are analogous to physical forces, i.e. "psychological forces" with an origin, direction and magnitude that is driving the causal agent (Wolff, 2007). Agents' intentions towards a certain state or goal can align or conflict, characterising the exact causal causal relationship between them ("The police officer enabled vs. prevented the woman from going to the other side of the road."). Wolff (2007)'s experiments demonstrate that people treat an agent's indication of their intention (e.g. pointing in the direction of where they want to go) analogous to physical force. In contrast to intentions or desires, knowledge states are less clearly goal or action-directed. Without further theoretical elaboration on how knowledge states can be modelled as psychological forces, we do not see how an integration of our findings on epistemic states and actions in this framework could work.

3.14 The Chicken or the Egg: Causality and Blame

The account we suggest here aims to provide a theory of how people make judgments about causality. We argue that agent epistemic states influence the perceived causality of an agent for an outcome. In particular, the theoretical proposal we suggested in this chapter is that people mentally build causal models that include an agent's epistemic states and epistemic actions, and counterfactually intervene on these epistemic variables in order to determine the agent's causality.

There is an active debate in both causal and moral psychology about whether people actually sharply distinguish between judgments about causation and judgements about blame or responsibility (Alicke & Rose, 2012a; Alicke et al., 2012; Danks, Rose, & Machery, in press; Samland et al., 2016b; Sytsma, 2020a, 2020b). While causality has traditionally been assumed to be assessed independent of and prior to aspects of blame and morality (Malle et al., 2014a; Malle & Knobe, 1997; Shaver & Drown, 1986), some have argued that the domain of causality and blame gets blended in people's responses about these matters. Judgements about causality are biased by attributions of blame (Alicke & Rose, 2012a) or used equivalently to responsibility in ordinary language (Danks et al., in press; Sytsma & Livengood, 2019). Others have raised the concern that a verbal test question about "a cause" or "causation" might be interpreted in a way to assess accountability (Samland & Waldmann, 2016), and will hence be influenced by factors that are relevant for an agent's accountability for the outcome. Notwithstanding whether causality blends with normality judgments on the cognitive level, or on the pragmatic level, the question arises whether the impact of epistemic states on causal judgments is influenced by normative judgments. Given the influence of our epistemic manipulations on judgements of blameworthiness for ignorance, such a concern is warranted.

3.14.1 The Link Between Counterfactuals and Blame

The account we aim to give here applies to causality of outcomes of different valences, and we have argued above why epistemic states – and in particular counterfactual interventions on epistemic states – should play a role for the perceived causality of agents bringing about unknown positive outcomes, too (Lagnado & Channon, 2008). The central assumption of our argument relies on the crucial role of counterfactuals in causal thinking, and in particular, the point of intervention in this process. Our experiments show that the point of change in counterfactual thinking about a causal agent is systematically influenced by the agent's epistemic conditions. Any account that postulates an influence of epistemic states on causal judgments via normative judgments will hence need to account for the normative influence on people's counterfactual responses as well. According to such a line of argument, both causal as well as counterfactual responses such as "If the doctor had known about the side effects ..." would be influenced by some sort of normative considerations, and/or used to express a blame response (Samland & Waldmann, 2015). Assuming that judgments of causation are in some way normatively influenced, the central question then becomes how these normative judgments factor into the counterfactual reasoning process.

3.14.1.1 Counterfactuals before Blame

One possible line of response is that the kind of counterfactuals people consider determine their attributions of blame (Sher, 2009). The role of alternative possibilities is a prominent one in theories of responsibility attributions (Widerker, 2017). In fact, explicit probing of counterfactual thinking has been shown to increase attributions of blame or responsibility (Branscombe, Wohl, Owen, Allison, & N'gbala, 2003; Mandel & Dhami, 2005; Sytsma, 2020b). Likewise, research by Markman and Tetlock (2000) demonstrates that the direction of influence between counterfactual thinking and normative judgment can go backwards: In order to defy attributions of blame, people counterfactually refer to the (im)mutability of epistemic states ("I couldn't have known that ..."). On such a view, people's causal judgements are a result of how the counterfactuals they imagine influence their attributions of blame and responsibility. Defenders of the normative account have in fact been open to the idea that counterfactual reasoning influences the responsibility judgements they take to affect judgements about causation. (see e.g. Sytsma, 2020b).

3.14.1.2 Blame before Counterfactuals

Alternatively, however, it might be argued that judgments about blame come even prior in the process and directly influence the counterfactuals we consider. On such an account, people intervene on epistemic states or variables because they identify these as targets of blame or express blame by doing so. The idea that moral evaluations influence counterfactual thinking which in turn affects judgments about non-moral properties is not new (Knobe, in press; Phillips et al., 2015, 2019). Studies have demonstrated that moral judgments about aspects in the actual world influence which kind counterfactuals people consider as most relevant (Hitchcock & Knobe, 2009; Icard et al., 2017; Kominsky & Phillips, 2019; Phillips et al., 2015). However, these theories have been silent about how people parse causal models and how to identify the variables that people aim to change or intervene on when considering the most "normal" counterfactual (Kominsky & Phillips, 2019). The account we have given has the advantage that it postulates reasons for epistemic interventions that are independent of the normality status of these epistemic states. Changing the doctor's epistemic state from ignorance to knowledge about the side effect might represent a morally good or statistically normal epistemic state, but we argue that people will intervene on epistemic states independent of how 'abnormal' the agent's state of ignorance is. As a result, we argue that interventions on ignorance drive people's causal judgements even if their state of ignorance is not blameworthy or morally bad. While the studies in the previous chapter have indeed shown that ignorance moderates the effect of normality, the results from the experiments in this chapter show that general ignorance about the consequences of one's action has an impact, too.

3.14.2 Pragmatic Effects: Conditioning on Actions

In a series of compelling studies, Samland and Waldmann (2016) aim to address pragmatic influences on verbal causal test question by employing alternative causal measures (Samland & Waldmann, 2015). In their experiments, they use conditional probability contrast measures, letting participants estimate the probability of the outcome given the absence or presence of the agent's action ("How likely is it that E occurred if S had acted/not acted"). They find that people's responses on these measures are not affected by an agent's norm violation. People judge the probability of the outcome to be similarly low given the non-action of a neutral vs. norm-violating agent. Given that such a measure might be less susceptible to considerations of blame, how would they work for our experiment? Crucially, we argue that the estimated probability of the outcome is sensitive to the kind of variable that is conditioned on. Conditioning on the agent's action cuts off any prior epistemic variables and will hence not be affected by epistemic states. We would predict people's responses on a measure such as "If Dr. Jones had not administered Afibo, how likely is it that the cramps would have occurred?" to be low for both a knowing and an ignorant causal agent. In contrast, when conditioning on knowledge, e.g. ("How likely is it that E occurred if S had not known/known about p?") we would expect a difference in the estimated probability of the outcome.

The discussion of adequate causal test measures, however, highlights another important aspect of our account. We predict that causal test measures that are narrowly focused on the agent's causal action ("Did the Doctor's prescribing of the drug cause the side effect?") will not be affected by the agent's epistemic states, since they direct people's counterfactual intervention on the action. Our account applies to "causal agents", broadly construed.

In sum, we are open to the idea that, broadly, causal judgments in our experiments might be directly or indirectly influenced by normative judgments. The main question then becomes at which point in the process of couterfactual reasoning blame judgments come in. Our central assumption, however, that counterfactual reasoning - and in particular, counterfactual reasoning over epistemic states - is the core mechanism that is underlying the formation of these judgements, remains.

3.15 Conclusion

Could the tragic death of Romeo and Juliet have been avoided? Dependence theories of causation have famously argued that our thinking about how things could have gone differently underpins our ability to identify causal relationships and the assessment of the causal strength of causal factors. While there are many things in the story of Romeo and Juliet that could or should have taken a different course, even in the hopeless final scenario, it seems that their death could still be avoided. Intuitively, if Romeo did not falsely believe that Juliet was dead, that is, if he knew that Juliet was still alive, he would not have poisoned himself, and in consequence, Juliet would not have stabbed herself out of grief. In causal scenarios that involve human agents, agent's knowledge states, foresight or even intentions pick out those causes that can be controlled or manipulated in order to achieve desirable future outcomes. In this chapter, we have argued that epistemic states and conditions play a crucial role for people's causal judgments about them. We have shown that epistemic states function as points of interventions in people's counterfactual reasoning about the causal scenario. Taking into account mental states might explain why we attribute lesser causality to ignorant agents, but also acknowledges the forward-looking dimension of causal judgments. In addition, it might allow us to bridge the conflict between two classes of theories – blame vs. causality-oriented – which have long been fighting over the prerogative of interpretation of people's causal judgments. While various studies in psychology have demonstrated the influence of agent mental states on causal reasoning about human agents, we envisage this chapter to provide a first answer to the question *why* this is the case.

Chapter 4

Inferences from Causal Explanations

4.1 Introduction

Imagine the following scenario: Your flatmate Suzy recently applied to medical school, and today she will find out whether she has been accepted. In order to be accepted into the medical programme, she needs to pass two entrance tests: a test on physiology, and a test on anatomy. You remember that Suzy told you that she knows a lot about one topic, but unfortunately knows very little about the other topic. However, you don't remember which topic she knows well, and which she doesn't. Later that day, you hear Suzy cheering from her room. When you enter the room to ask her what happened, she replies: "I got into med school because I passed anatomy!". Is anatomy the topic that Suzy knows well, and was therefore likely to pass? Or is it the topic she knew poorly, and was unlikely to pass?

From what Suzy said, we not only know that she got into med school and that she passed anatomy. We also know that these events were causally related – passing anatomy helped get Suzy into med school. Do we learn anything more about what happened? Intuitively, it seems more likely that anatomy was the topic that Suzy did not know much about. This example demonstrates how a causal explanation can be informative about features beyond what was explicitly stated. In this case, the listener learns that Suzy's passing anatomy was unexpected. How is it that we manage to learn so much from such minimal input?

4.1.1 Aim of this chapter

This question of how we learn from other people's causal explanations is especially acute since, on the face of it, we seem to say very little explicitly when offering explanations (cf. Hemmatian & Sloman, 2018, for a striking example, where mere labels are used as explanations). A prototypical causal explanation may involve nothing more than a specification that "E happened because C happened," essentially just citing two events, C and E. As Keil (2006) frames the issue, "Somehow, people manage to get by with highly incomplete or partial explanations of how the world around them works [...] We have yet to understand the nature of such compressions of information" (p. 135).

The aim in this final chapter is to establish groundwork for the subtle but systematic patterns of inference people draw upon receiving an explanation. The idea that listeners in a conversation go far beyond what is explicitly (or "literally") said by a speaker is widely appreciated, and there are well-developed theoretical frameworks for studying this capacity (e.g., H. H. Clark, 1996; Goodman & Frank, 2016; Levinson, 2000). In this chapter, we aim to offer an account of how inferences from causal explanations might work. How do we manage to learn so much from so little?

As a way into this question, we want to leverage again on the growing body of research on how normality and causal structure affect people's causal judgments (e.g., Gerstenberg & Icard, 2019; Icard et al., 2017; Kominsky et al., 2015). As discussed in previous chapters, experimental work has emphasised systematic patterns in participants' judgments about various causal claims. In the present studies, we will turn this around, probing not just what causal explanations people deem reasonable, but whether they can leverage these very intuitions to make appropriate inferences from claims made by others. Specifically, we show in two studies that people are able to infer specific information about *normality* when provided information about *causal structure*, and conversely, they can infer causal structure when provided with information about event normality. These case studies, we argue, underscore much of what is characteristic of inferences from causal explanations.

People's inferences stem from a combination of commonly shared causal intuitions, conversational principles, and finally as well leveraging on their fundamental ability of perspective taking.

4.1.2 Learning from Explanations

Much of what we know about the causal structure of the world we infer from directly observing and interacting with it (Cheng, 1997; Cheng & Novick, 1990; Gopnik et al., 2004; Lagnado et al., 2007; Waldmann & Hagmayer, 2001). We also observe others take actions, and learn from their successes and failures (Bandura, 1962; Bekkering, Wohlschlager, & Gattis, 2000; Hanna & Meltzoff, 1993; Jara-Ettinger et al., 2016; Whiten, 2002). The way we learn about the world from explanations – from utterances of the form "*E* because *C*" – has its own distinctive character (Hesslow, 1988; Hilton, 1990; Turnbull & Slugoski, 1988). Rather than observing or experiencing a sequence of events directly, we receive a kind of packaged summary of the relevant events; and, if successful, this summary allows us to make appropriate inferences about important aspects of the situation. As a result, learning from explanations often involves communication.

4.1.3 Explanations in Communication and the Role of ToM

Generically, communication involves (at least) two interlocutors with partially overlapping knowledge about the world (H. H. Clark, 1996). Part of what makes communication such an efficient information transfer possible is the fact that people generally adhere to cooperative communicative principles in how they produce and interpret linguistic utterances (H. P. Grice, 1975). That is, we can generally rely on people to be as informative as they can, to be truthful, to mention only relevant things, etc. This allows the meanings of utterances to be relatively underspecified, since both speaker and listener can rely on a combination of world knowledge and tacit understanding of conversational principles to go far beyond what is said literally (Goodman & Frank, 2016; H. P. Grice, 1975; Levinson, 2000). The ability to understand someone's mental states and and how these are connected with behaviour is an undisputed requirement of human communication (Happé & Loth, 2002; Sperber & Wilson, 2002). In particular, pragmatic competence, i.e. the functional use and interpretation of language in social contexts requires the use of both linguistic and extra-linguistic communication in context, such as the ability to attribute mental states to others. The ability to understand another speaker's intended meaning has been argued to overlap, or even be dependent on Theory of Mind (Bosco, Tirassa, & Gabbatore, 2018; Fernández, 2013; Zufferey, 2010). Pragmatic communication is possible to the extent that both speaker and listener are able to entertain thoughts about the mental states of the other. In a communicative exchange, they attribute a rich array of beliefs, desires and knowledge to each other's minds in an effort to recover the particular communicative intention that had motivated the speaker's utterance. A listener uses their knowledge about the speaker's mental states to derive the implicature of their utterance. On the other hand, the speaker is able to issue such an implicature because they know certain things about the listener's mental states. As such, pragmatic communication is a core element of social interaction that, to some extent, hinges on a well-developed ToM (Cummings, 2015a, 2015b). In fact, pragmatic language skill and mental state understanding are reciprocally interconnected at different points throughout children's development (Westra & Carruthers, 2017).

Pragmatic comprehension crucially involves understanding people's minds. But how does it enable people to make inferences from explanation beyond what is explicitly said? Our proposal draws on two aspects: some simple but general principles of communication, and a minimal analysis of what the signal "*E* because *C*" means. For a first pass at the latter, we take the meaning to be captured by the circumstances in which it would be appropriate for the speaker to utter the phrase. These two ingredients then allow us to predict what people will infer from a causal explanation. If a speaker *S* utters to a listener *L*, "*E* because *C*," then *L* may think about how the world must have been in order for this to have been an appropriate thing for *S* to say. Assuming that *S* is a cooperative speaker, using the phrase in the normal way, and knowledgeable about the relevant state of the world, *L* will be able to infer a certain state of the world by simulating what the speaker would have

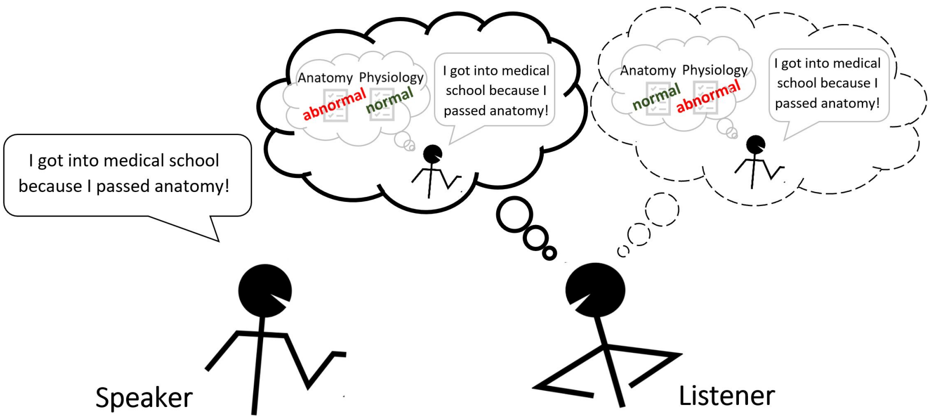

I got into medical school because I passed anatomy!
Anatomy Physiology
abnormal normal
I got into medical school because I passed anatomy!
Anatomy Physiology
normal abnormal
I got into medical school because I passed anatomy!
Speaker
Listener

Figure 4.1: Illustration of the communicative situation of the "Suzy" example. The listener knows that Suzy needs to pass both Anatomy and Physiology in order to get into medical school. The listener doesn't know which one she is more likely to pass. Upon hearing the speaker's explanation, the listener considers what they would have said in each of the two possible situations. Because speakers have a tendency to refer to abnormal events in their causal explanations, the listener infers that anatomy was the subject Suzy was less likely to pass.

found most reasonable to say.

To illustrate, consider again our running example (see Figure 4.1). Suzy's utterance is consistent with two possible states of the world. As listeners, we know that acceptance to medical school requires passing both physiology and anatomy, and that Suzy is unlikely to pass one of them, but we don't know which one. The statement "I got into medical school because I passed anatomy" prompts us to consider two possible scenarios in which Suzy might have made this statement: the scenario in which anatomy was the subject that Suzy was unlikely to pass, and the scenario in which it was physiology. Evidently, we have a strong preference for the situation in which the cited cause represents the abnormal event. Why? Intuitively, only in this scenario would Suzy's utterance have been a sensible thing to say. [1]

Crucially, this process relies on a minimal reference to the speaker's mental

[1] In this example, Suzy not only utters an explanation, but she seemingly does so with an emotive undertone: She is happy that she was accepted into medical school. This might raise the question to what extent people's inferences are driven by the affective component of her explanation. In this chapter, we focus on investigating people's inferences from explanations that are presented in a neutral manner (selected verbal statements), without further information about the speaker's attitudes towards causes and outcome. While this example might suggest otherwise, we will show that systematic inferences do not require an explanation to be affect-laden.

states, and what they would deem as an appropriate utterance. If the explanation "Suzy got into medical school because of anatomy" would have been generated by a machine, or the reference to anatomy would have been selected by a random coin flip, no such systematic inference about the abnormality of the anatomy test would be possible. It is the reference to the speaker's own explanatory preferences that allows us to make these inferences.

The fact that citing anatomy as the cause strikes us as sensible is just one instance of the well-known trend whereby people prefer to cite abnormal or unexpected events as causes (Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Kahneman & Miller, 1986). Indeed, there is a wealth of existing experimental work on the factors that influence what causal explanations people judge appropriate. Though there is a comparative paucity of work on what inferences people draw from others' explanations, we argue that all of this existing work provides a useful starting point, once embedded in a suitable communication-theoretic framework. In short, if listeners know that speakers' explanations follow systematic patterns, they should be able to infer what happened simply by considering what would have been reasonable to say across the relevant possible states of the world. In this chapter we focus again on the two especially well-studied factors that are known to shape causal judgements explanations: norms and causal structure.

4.1.4 The Influence of Norms and Causal Structure on Causal Explanations

When multiple causes contributed to an outcome, people tend to select a few causes in their explanation of what happened rather than citing all of them. Causal selection moreover follows systematic patterns. As discussed in previous chapters, people often prefer abnormal over normal events as causes (Cheng & Novick, 1991; Halpern & Hitchcock, 2015; Hart & Honoré, 1959/1985; Hesslow, 1988; Hilton & Slugoski, 1986; Phillips & Cushman, 2017). When two causes are each necessary for producing a certain outcome (conjunctive structure), people judge the abnormal event as more causal (Gerstenberg & Icard, 2019; Icard et al., 2017; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Kominsky et al., 2015). The influence of normality on causal selections has been shown both for *statistical norms* (i.e. the frequency with which an event occurred in the past), as well as *prescriptive norms* (i.e. whether an event adheres to or violates a social or moral norm).

While there is an ongoing discussion on how to best explain this preference for abnormal causes (Alicke, 2000; Kominsky & Phillips, 2019; Samland & Waldmann, 2016; Sytsma et al., 2012), recent research has found that when two causes are each sufficient for the outcome (disjunctive structure), people show a preference for the *normal* over the abnormal cause (Gerstenberg & Icard, 2019; Henne, Niemi, et al., 2019; Icard et al., 2017). Coming back to our example from the beginning, imagine a disjunctive test situation in which passing either anatomy or physiology (or both) is sufficient for getting accepted into medical school. If Suzy passes both exams in this situation, people should be more likely to explain her acceptance by referring to the test that she was expected to pass (i.e. the "normal" cause). This effect is surprising because unlike what has been assumed for decades (e.g. Hart & Honoré, 1959/1985), people don't show a uniform preference for abnormal causes. Instead, event normality and causal structure interact to determine causal selections.

Why do people perceive a normal factor as more causal when the causal structure is disjunctive? Icard et al. (2017) suggest that the perceived causal strength of a cause is a function of its necessity and sufficiency weighted by the normality of the event. Others have argued that the correspondence in normality between cause and effect is what matters for causal selections (Gavanski & Wells, 1989; Harinen, 2017). People select abnormal causes for abnormal effects, and normal causes for normal effects. Currently, there are a number of competing hypotheses about *how* causal structure and normality affect causal selections, but we still lack a complete understanding for *why* they do (cf. Fazelpour, 2020).

Research into the effects of normality on causal judgments has predominantly used written vignettes involving norm-violating agents. The fact that these vignettes often involve intentional human agents has prompted some to argue that, rather than assessing causal judgments per se, people's responses may be shaped by concerns of accountability or blameworthiness (Alicke, 2000; Samland & Waldmann, 2016;

Sytsma et al., 2012). Verbal descriptions of causal scenarios also leave some uncertainty about the causal structure, and how much each agent actually contributed to the outcome. In addition, it has been argued that the effects of normality might be restricted to language-driven forms of causal thinking (Danks, Rose, & Machery, 2014).

A recent study by Gerstenberg and Icard (2019) suggests that the effect of norms on causal cognition are more far-reaching than previously thought. In their experiments, participants watched video clips showing physically realistic interactions between inanimate objects (see Figure 4.2). In these clips, ball A and ball B enter the scene from the right, and are headed toward a stationary ball E. In order to hit ball E, each of them needs to pass through a blocker. Crucially, the blockers differ in how likely they are to let a ball go through. While the light red blocker has a 80% chance of letting a ball go through, the dark red blocker only has a 20% chance. The clips came in two different setups that were manipulated between participants: In the conjunctive setup, both ball A and ball B need to go through the blocker in order to make ball E through the gate. In the disjunctive setup, being hit by either ball A or ball B is sufficient to make ball E go through the gate. Participants watched ten of these clips and learned how likely it was for each blocker to let a ball go through. In the test phase, participants watched a clip in which both balls went through the blocker and, as a result, ball E went through the gate (Figure 4.2 middle). Participants were asked to select which explanation better described what happened: "Ball E went through the gate because ball A [ball B] went through the motion block".

The results showed – consistent with prior research – that when two causal events were both necessary to make the outcome happen, participants selected the abnormal event (i.e. the ball that was unlikely to go through the blocker). However, when either of two events was individually sufficient, participants selected the normal event (i.e. the ball that was likely to go through the blocker). Statistical norms affect causal selections even when there is little to no uncertainty about what actually happened (unlike in vignette studies), and when the setting is purely physical

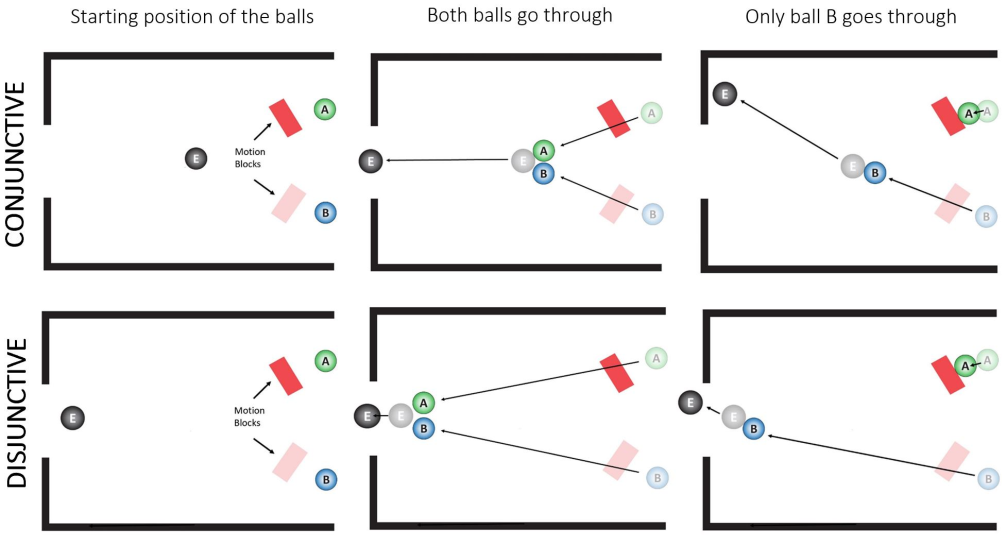


Figure 4.2: Diagrammatic illustration of clips used in Gerstenberg and Icard (2019) (original clips varied slightly). The top row shows the conjunctive causal structure, the bottom row the disjunctive structure. The color of each blocker indicates its probability of blocking a ball. The dark red blocker has an 80% chance of blocking a ball, and the light red blocker has a 20% chance of blocking a ball. The first column shows the starting position of the balls, the second a case in which both balls went through the blocker, and the third column a case in which only ball B went through the blocker.

so that potential effects of accountability or blameworthiness are minimized (if not absent).

4.2 Predictions

To summarise, we propose that the inferences people draw from others' explanations can be predicted on the basis of general principles of conversation together with an accurate construal of what people take claims of the form "*E* because *C*" to mean. Just like other types of utterances, explanations follow general discourse rules. Explanations are necessarily incomplete and speakers can rely on listeners to fill out the missing gaps, on the basis of a common understanding about the world and shared general principles of communication. What a listener learns from an explanation of the form "*E* because *C*" goes beyond what is explicitly stated. Listeners can infer information about the contextual factors that would have made this explanation an appropriate thing to say. Knowing the general regularities that influence people's selection of causal explanations enables us to predict the kind of information that is inferred from an (incomplete) causal explanation.

We assume the large body of research on causal judgment – including on the roles of norms and causal structure – offers a suitable hypothesis about the latter. This broad proposal issues in a number of concrete predictions, which we outline below.

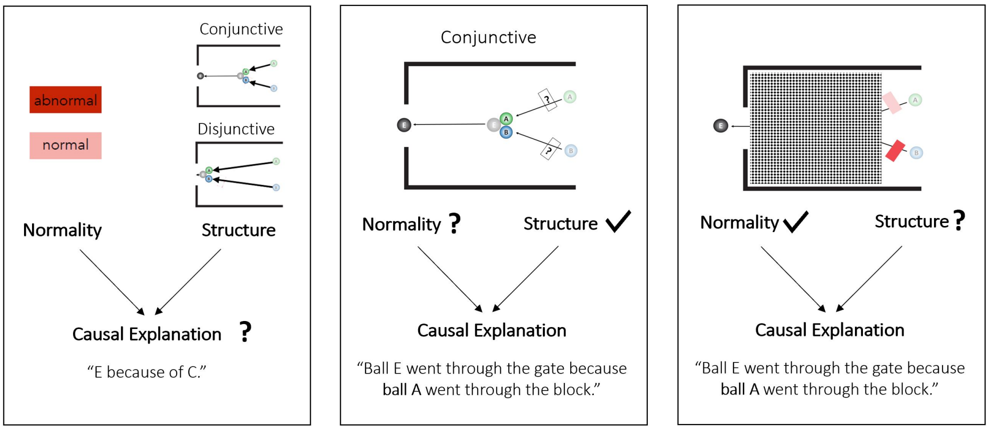
4.2.1 Hypotheses

4.2.1.1 Hypothesis 1 (Replication): People's selections of causal explanations are influenced by event normality and causal structure

From prior research, we know that both the normality of causes as well as the underlying causal structure influence causal judgments and explanations (Gerstenberg & Icard, 2019; Icard et al., 2017; Kominsky et al., 2015). As a first hypothesis, we predict a replication of these effects in our experiments. Specifically, we predict that when an abnormal and a normal cause bring about an outcome E, people will tend to select the abnormal cause as an explanation for why E happened when both causes were needed ("conjunctive causal structure"), but will tend to select the normal cause when either cause would have been sufficient ("disjunctive causal structure"; Figure 4.3a). While most prior work has found these kinds of effects on continuous causal judgments (e.g. Icard et al., 2017; Kominsky et al., 2015), we predict that the same pattern will hold for discrete causal selections (see also Gerstenberg & Icard, 2019). Furthermore, we wanted to replicate these effects because most prior work has used written vignettes whereas in our experiments we show participants animated causal scenarios.

4.2.1.2 Hypothesis 2: People infer an event's normality from an explanation given knowledge about the causal structure

When a causal explanation is given and the causal structure is known, we predict that people can infer the corresponding normality of the cited cause (see Fig-



Hypothesis 1: Influence of normality and structure on causal explanations.

Hypothesis 2: Inference about normality from causal explanation and structure.

Hypothesis 3: Inference about structure from causal explanation and normality.

Figure 4.3: Diagrammatic illustration of Hypotheses 1, 2 and 3. Hypothesis 1 predicts that both normality as well as causal structure determine which event people cite as a cause of ball E's going through the gate. Hypothesis 2 predicts that people can infer the normality of the blockers based on the underlying causal structure and a given causal explanation. Hypothesis 3 predicts that people can infer the underlying causal structure of the scene based on the normality of the blockers and a given causal explanation.

ure 4.3b). More precisely, knowing that there are only two possible options, they can infer whether the cited cause was abnormal or normal. We assume that people make this inference by considering what they themselves would have said in the given situation. For instance, consider the example in Figure 4.1. Here, the listener knows that the structure is conjunctive – the speaker needed to pass both anatomy and physiology to get into medical school. When the speaker states that she got into medical school because she passed anatomy, the listener infers the event normality by considering how likely she would have said the same thing in the two possible situations. Given the general preference for citing abnormal causes in conjunctive situations, the listener infers that passing anatomy was abnormal. A speaker would be more likely to cite passing anatomy as the cause when this event was abnormal compared to when it was normal. More generally, if the causal structure is conjunctive, participants will infer that the cited cause is likely to be the abnormal event. In contrast, if the causal structure is disjunctive, participants will infer that the cited cause is likely to be the normal event.

4.2.1.3 Hypothesis 3: People infer the causal structure from an explanation given knowledge about the event's normality

We predict that people can infer the causal structure of the situation based on what they know about the normality of the cause cited in the explanation (see Figure 4.3c). Again, this prediction rests on the assumption that people make this inference considering two concrete hypotheses according to which the causal structure is either conjunctive, or disjunctive.

For example, consider a situation in which the abnormal event is cited as the cause. In this case, the listener has to ponder what cause they would cite if the causal structure was conjunctive, and what cause they would cite if the structure was disjunctive. The listener also has to take into account the prior probability of the structure being conjunctive or disjunctive. Given that we know that people generally have a preference for selecting the abnormal event for conjunctive structures, and the normal event for disjunctive structures, we predict that most participants will infer a conjunctive structure if the abnormal event is cited as the cause, and a disjunctive structure if the normal event is cited.

4.2.1.4 Hypothesis 4: Individual differences in inferences from explanations

In the general case, we expect that listeners make use of their ToM: They take into account their knowledge of the speaker and the speaker's own preferences when interpreting the speaker's explanations, (Goodman & Stuhlmüller, 2013; Kamide, 2012; Schuster & Degen, 2019; Yildirim, Degen, Tanenhaus, & Jaeger, 2016). For example, if a listener happened to know that a speaker has a general tendency to cite abnormal events as causes no matter what the causal structure is, then the listener wouldn't be able to infer the causal structure when the speaker cited an abnormal cause. In the settings that we consider, listeners don't have any speaker-specific information. Accordingly, we assume that listeners will consider what explanation they themselves would have given. That is, in the absence of knowing the speaker's particular preferences and likes, people refer to their own "mind", i.e. their own preferences and desires, in order to draw inferences.

In our experiments, we ask participants to select explanations themselves (Hypothesis 1), and to infer what happened from hearing another person's explanation (Hypotheses 2 and 3). We predict that there will be a close correspondence between individual participants' explanation preferences and their inferences. For example, we expect that a participant who selects an abnormal cause in a conjunctive situation, and a normal cause in a disjunctive situation, will be more certain about the underlying causal structure upon hearing an explanation that cites an abnormal cause, compared to a participant who has a general preference for selecting abnormal causes. We will spell out these predictions about how individual differences affect inferences from explanations in more detail in the results section of each experiment.

In the following, we will report two experiments testing these hypotheses. We test Hypothesis 1 in both experiments. Additionally, Experiment 1 tests Hypothesis 2 and Experiment 2 tests Hypothesis 3. For both experiments, we will look at aggregate results, but also analyse the data by taking into account interindividual differences. Both experiments use two types of norm violations: *statistical norm* violations (using the billiard ball setup shown in Figure 4.2), and *prescriptive norm* violations involving a scenario with intentionally acting agents (Figure 4.4).

4.3 Experiment 1: Inferring Normality given Causal Structure

In Experiment 1, we test whether participants can infer the normality of an event cited in an explanation based on knowledge about the causal structure of the situation.[2]

[2] All the materials including data, figures, videos, and analysis scripts may be accessed here:`https://github.com/cicl-stanford/inference_from_explanation`

4.3.1 Methods

4.3.1.1 Participants and Design

We recruited 210 participants (Mean$_{\text{age}}$ = 33, SD$_{\text{age}}$ = 9, N_{female} = 77, $N_{\text{non-binary}}$ = 2, $N_{\text{undisclosed}}$ = 4) via Amazon Mechanical Turk (Crump, McDonnell, & Gureckis, 2013). 56 participants were excluded for failing one or more exclusion criteria specified below, leaving a final sample size of $N = 154$ (26.7% excluded). The experiment has a 2 causal structure (conjunctive vs. disjunctive) × 2 norm type (statistical vs. prescriptive) design. Both factors were manipulated between participants. Participants were randomly assigned to the four separate conditions, *statistical normality & conjunctive structure* ($N = 30$), *statistical normality & disjunctive structure* ($N = 37$), *prescriptive normality & conjunctive structure* ($N = 46$), and *prescriptive normality & disjunctive structure* ($N = 41$).

4.3.1.2 Statistical Normality: Selection Task

We closely followed the paradigm in Gerstenberg and Icard (2019). Participants were informed that they were going to see video clips of colliding billiard balls, followed by a diagram and description of the billiard ball setup (see Figure 4.2).

In the *conjunctive* condition, participants saw a diagram illustrating that both balls A and B needed to go through the blockers in order for ball E to go through the gate (see Figure 4.2, "Conjunctive"). In the *disjunctive* condition, participants, were informed that either ball A or ball B's going through the blockers is sufficient for ball E to go through the gate (see Figure 4.2, "Disjunctive").

In our experiment, participants were informed that the position of the two blockers may vary from scene to scene. In some setups, the light red blocker would be at the top, while in others, the light red blocker would be at the bottom. Subsequently, participants were asked a series of comprehension check questions about the billiard ball setup. For example, participants were shown a diagram of a situation and then asked: "In this set-up, if only one of the balls go through the motion block and hit ball E, ball E will go through the gate" with the response options being true/false. See the materials posted online for the full list of comprehension check questions: `https://github.com/cicl-stanford/inference_from_explanation`. In order to be included in the experiment, participants had to answer all check questions correctly, and they were given three attempts to read through the instructions and answer the check questions correctly. If a participant failed to answer the check questions after the third attempt, they were forwarded to the end of the study and thanked for their participation.

Participants who answered all of the comprehension check questions correctly then continued with a task in which they themselves selected a causal explanation. This task served two purposes. First, to further familiarise participants with the scenario. Second, to acquire data on participants' own explanation preferences. Participants first only viewed the beginning of the clip. The clip paused shortly after ball A and B entered the scene. Participants were asked to what extent they agreed with the following three predictions: (1) "Ball A will hit ball E.", (2) "Ball B will hit ball E.", and (3) "If only one of the two balls goes through the block and hits ball E then ball E will go through the gate." Participants provided their responses on sliding scales with the end points labeled as "not at all" (0) and "very much" (100). The position of the blockers was counterbalanced across participants. We only included participants in the final analysis who rated the chance of the normal billiard ball to hit Ball E higher than that of the abnormal ball, and who responded < 50 in the conjunctive or > 50 in the disjunctive condition for statement (3). These attention check questions made sure that participants had correctly encoded the information in the instructions. The clip then continued to play. Both balls went through the blocker and ball E went through the gate. Participants were asked to select which of the following two statements better described what happened: "Ball E went through the gate because ball A / ball B went through the motion block." We used a two-alternative forced-choice task (rather than a continuous judgment) to match the explanation format that participants received later in the test phase.

4.3.1.3 Statistical Normality: Inference Task

In the final inference task, participants received a diagram showing a conjunctive (or disjunctive) billiard ball setup in which both ball A and ball B went through the blocker and ball E went through the gate. However, the causal diagram did not include any information about the normality of the two blockers (see Figure 4.3b). Participants were then told that Ben, a fictional participant, had witnessed the depicted scene and, as in the selection task before, had been asked to select an explanation that best explained what happened. Participants were told that Ben selected the explanation "Ball E went through the gate because ball A [ball B] went through the blocker." We counterbalanced which ball Ben's explanation referred to (ball A or ball B).

Finally, participants were asked to indicate which scenario they thought Ben had seen. More precisely, they had to indicate whether Ben saw a scenario in which the ball he selected was likely or unlikely to go through the blocker. Hence, this task creates a minimally communicative situation in which the participant acts as a listener who receives a speaker's explanation. Participants indicated their response on a slider which showed one of two possible versions of the scenario at each end point of the scale. For example, on the left side of the slider they saw a scenario in which the unlikely dark red blocker was at the top and the likely blocker was at the bottom, and on right side a scenario in which the light red blocker was at the top and the dark red blocker at the bottom. Both endpoints of the slider were labeled "Definitely this one", referring to the scenario depicted above the endpoint. For example, if Ben chose ball A and ball A went through the top blocker, participants could indicate that this ball was an unlikely cause by sliding to the left, or that it was a likely cause by sliding to the right. The mid-point of the scale was labelled "Unsure". We counterbalanced which normality version of the scenario was shown on the left and which one on the right.

4.3.1.4 Prescriptive Normality: Selection Task

To manipulate prescriptive normality, we created an animated video version of the "motion detector" vignette from Kominsky et al. (2015). In this vignette, Suzy and Billy work on a project of national security and they both share an office. This office has a motion detector. In the *conjunctive* condition, the motion detector goes off if more than one person enters the office (Figure 4.4a, Conjunctive). In the

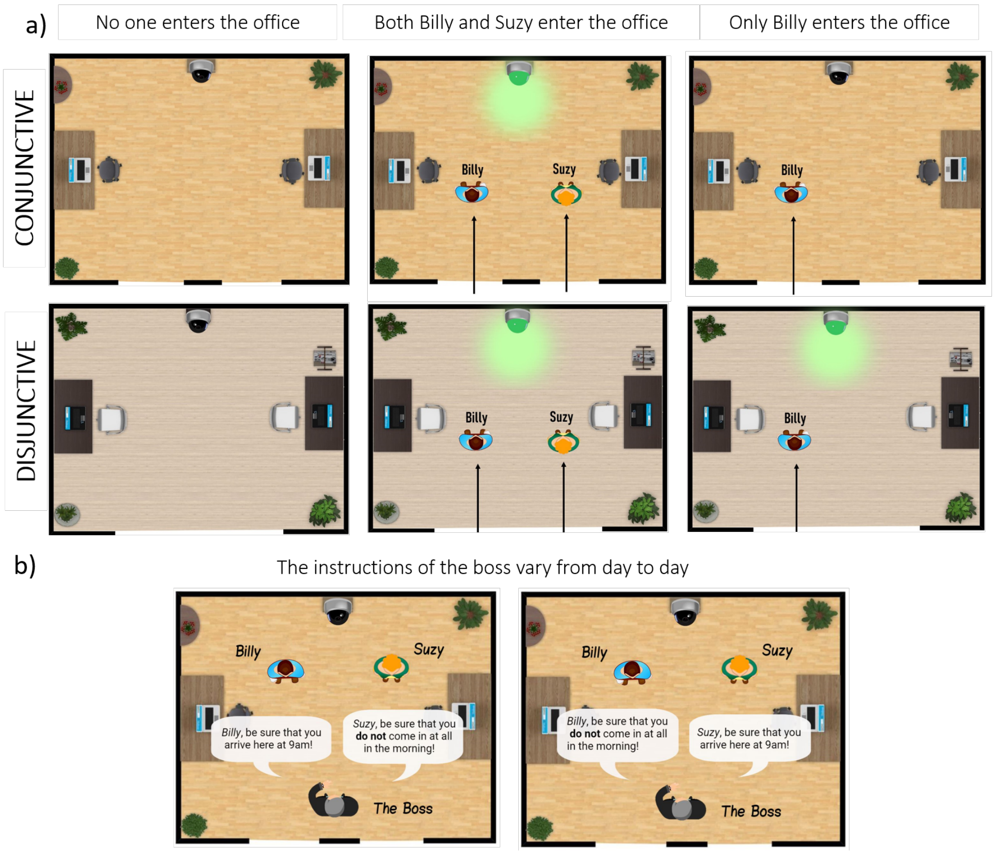

a)
No one enters the office
Both Billy and Suzy enter the office
Only Billy enters the office
CONJUNCTIVE
DISJUNCTIVE
Billy
Suzy
Billy
Billy
Suzy
Billy
b)
The instructions of the boss vary from day to day
Billy
Suzy
Billy, be sure that you arrive here at 9am!
Suzy, be sure that you do not come in at all in the morning!
The Boss
Billy
Suzy
Billy, be sure that you do not come in at all in the morning!
Suzy, be sure that you arrive here at 9am!
The Boss

Figure 4.4: a) Diagrammatic illustration of the animated clips of the "motion detector" vignette (cf. Kominsky et al., 2015). The top row shows the office including a motion detector with conjunctive causal structure, the bottom row the motion detector with disjunctive causal structure. The first column shows what happens when no one enters the office, the second a case in which both Billy and Suzy enter the office, and the third column a case in which only Billy enters the office. b) The instructions of the boss to the employees vary from day to day.

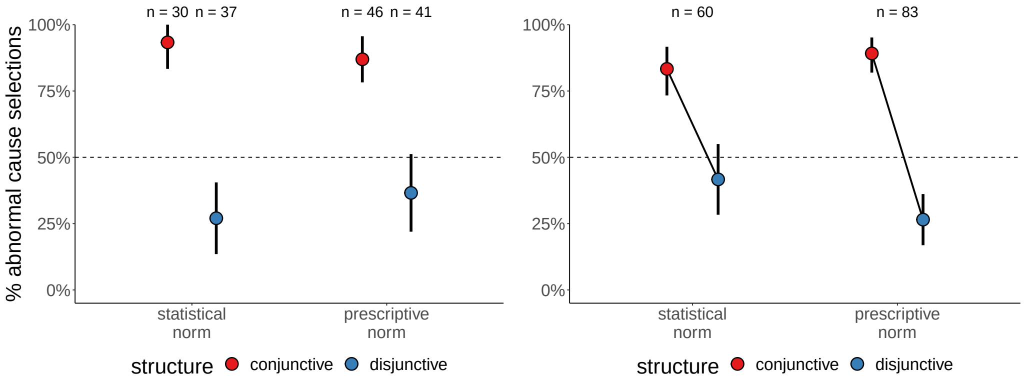
disjunctive condition, the motion detector goes off as soon as one person enters the office (Figure 4.4a, Disjunctive).

Given the confidentiality of the project, it is sometimes required that one employee works alone in the office. As a result, on certain days the company's boss instructs both employees that either only Suzy or Billy should come into the office at 9am the next morning, while the other one is supposed to stay away from the office (Figure 4.4 b). Who is instructed to come in and who to stay away may vary from day to day. Participants were provided with written instructions and the diagrams in Figure 4.4, followed by four comprehension questions that they needed to answer correctly before proceeding.

In both conditions, participants then saw a video, separated into two parts, showing one morning in the conjunctive [disjunctive] office, and what happened the day before. In the first part, the boss gives instructions to Billy and Suzy the day before. One of the two employees is told to arrive at 9am in the office the next morning, and the other is told to not come in during that time. We counterbalanced across participants who was the employee instructed to come in, and who to stay away. Participants were asked to what extent they agreed with the following three predictions: (1) “Billy is allowed to come into the office at 9am the next morning”, (2) “Suzy is allowed to come into the office at 9am the next morning”, and (3) “If only one of the two employees enters the office, the motion detector will go off.” Participants provided their responses on sliding scales with the end points labeled as “not at all” (0) and “very much” (100). We only included participants in the final analysis who responded > 50 for the normal agent, < 50 for the abnormal agent, and < 50 in the conjunctive or > 50 in the disjunctive condition for statement (3). The second part of the video showed the next morning. On this morning, both Suzy and Billy come into the conjunctive [disjunctive] office at 9am and, as a result, the motion detector goes off. Participants were asked to select which of the following two statements better described the scene: “The motion detector went off because Billy/Suzy entered the office.”

4.3.1.5 Prescriptive Normality: Inference Task

In the final inference task, participants received a diagram showing the office with the motion detector with the conjunctive [disjunctive] structure. On that morning, both Suzy and Billy entered the office at 9am, and the motion detector went off. However, the picture did not show what instructions the boss gave for that particular day (i.e. whether Billy or Suzy was supposed to come in at 9am). Participants were then told that Ben, a fictional participant, had witnessed the entire scenario including the day before when the boss gave the instructions. Ben was asked to select an explanation that best explains the observed scenario. Ben selected the explanation “The motion detector went off because Billy [Suzy] entered the office.” We counterbalanced which person Ben’s explanation referred to across participants.



(a) Experiment 1

(b) Experiment 2

Figure 4.5: Participant's causal selections in Experiment 1 and Experiment 2. The data points show the percentage of participants selecting the abnormal cause as a function of causal structure (conjunctive or disjunctive) and norm type (statistical or prescriptive). In Experiment 2, each participant made a choice for both structures as indicated by the lines connecting the data points. The causal selection task replicates and extends the pattern of causal selections from previous studies (Gerstenberg & Icard, 2019). *Note*: Error bars are bootstrapped 95% confidence intervals.

Participants were then presented with the question "Given Ben's decision, which of these two scenarios did he see?" They indicated their response on a slider with the two possible scenarios presented next to the slider endpoints. For example, an image of the scenario in which the boss instructs Billy to come in at 9am the next morning and Suzy to stay away would be shown on the left side, and the scenario in which Suzy is instructed to come in at 9am the next morning and Billy to stay away on the right side. Both endpoints of the slider were labeled "Definitely this one." referring to the scenario depicted above the slider end, and the midpoint was labeled "Unsure". Which scenario was depicted left and right was counterbalanced across participants.

4.3.2 Results

Figure 4.5a shows participants' causal selections as a function of the causal structure of the scenario (conjunctive vs. disjunctive) and the type of norm that was manipulated (statistical vs. prescriptive). Table 4.1 shows the results of an ANOVA on a series of generalised linear models. These results show that there was a significant effect of structure on causal selections $\chi^2(1) = 58.77, p < .001, \phi_c = 0.62$, 95% CI $[0.46, 0.78]$ (see Table 4.1). Participants were more likely to select the abnormal cause for conjunctive causal structures (89%) compared to disjunctive structures (32%). There was no effect of type of norm on participants' selections ($p = .76$), and no interaction effect between structure and norm ($p = .21$).

Causal structure also strongly affected participants' inference judgments (see Table 4.2), $F(1,150) = 66.59, p < .001, \eta_p^2 = 0.32$, 95% CI $[0.2, 0.43]$. Participants inferred that the event cited in the explanation was abnormal when the causal structure was conjunctive ($\mathrm{M} = 84.58, \mathrm{SD} = 28.13$), and normal when the structure was disjunctive ($\mathrm{M} = 41.42, \mathrm{SD} = 34.99$). Note that people were more certain that the cited cause was abnormal in the conjunctive structure than that it was normal in the disjunctive structure. The norm manipulation (statistical vs. prescriptive) had no effect on participants' inferences, and there was also no interaction effect between structure and norm type. In both selection and inference cases, the effect sizes can be considered large with $\phi_c > 0.5$ for the causal selections, and $\eta_p^2 > 0.25$ for the normality inferences

Table 4.1: Experiment 1 – Causal selection: **Experiment 1**: Analysis of variance of the causal selection results.

model specification: `selection ~ structure * norm`

Term	LR Chisq	Df	Pr(>Chisq)	Cramers_v	CI	CI_low	CI_high
structure	58.77	1.00	0.00	0.62	0.95	0.46	0.78
norm	0.09	1.00	0.76	0.02	0.95	0.00	0.17
structure:norm	1.57	1.00	0.21	0.10	0.95	0.00	0.26

Table 4.2: Experiment 1 – Normality inference: Analysis of deviance of the causal selection results.

model specification: `normality rating ~ 1 + structure * norm`

Term	Sum Sq	Df	F value	Pr(> F)
structure	67911.03	1.00	66.59	0.00
norm	332.89	1.00	0.33	0.57
structure:norm	348.55	1.00	0.34	0.56
Residuals	152966.72	150.00		

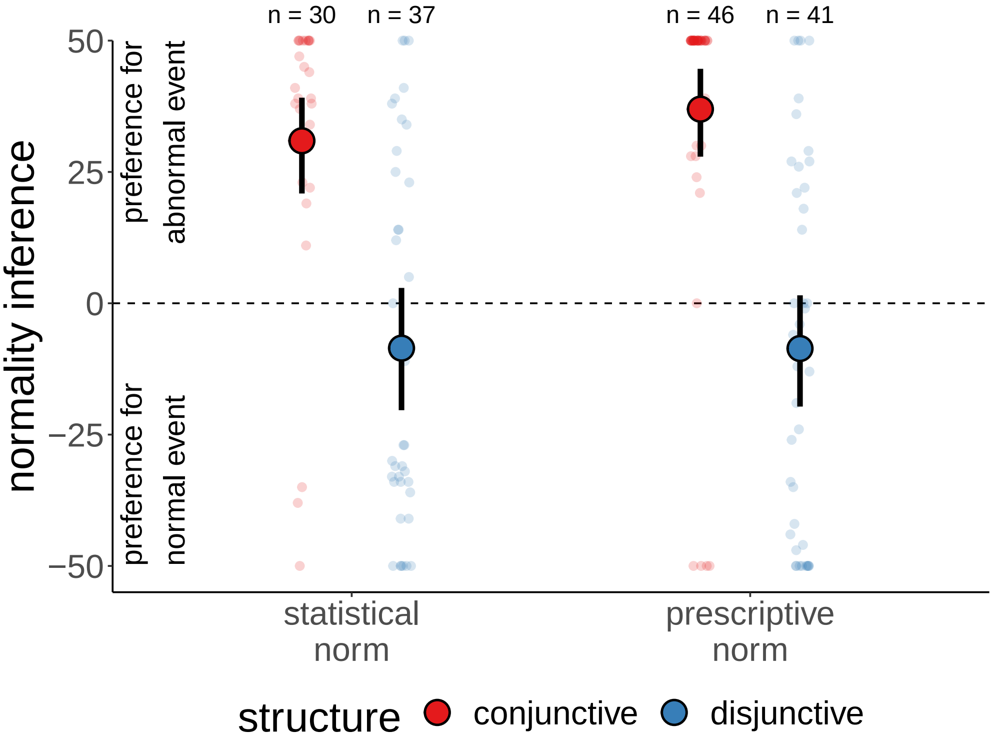


Figure 4.6: Experiment 1: Participants' preference for the abnormal cause (top) versus the normal cause (bottom) as a function of the causal structure (conjunctive vs. disjunctive) and the norm type (statistical vs. prescriptive). For example, the red data points show that participants who received an explanation for a conjunctive causal structure tended to infer that the cited cause was abnormal. *Note*: Large circles are group means. Error bars are bootstrapped 95% confidence intervals. Small circles are individual participants' judgments (jittered along the x-axis for visibility).

As hypothesised, we found a close correspondence between participants' causal selections and the inferences they drew about the normality of a cause given knowledge about the causal structure. This correspondence becomes even clearer

Table 4.3: Experiment 1 – Relationship between selections and inferences: Percentage of participants who selected the abnormal cause (selection), and who had a preference for inferring the abnormal cause (inference) as a function of the norm type and the causal structure.

norm type	causal structure	% abnormal cause selection	 inference
statistical	conjunctive	93	90
statistical	disjunctive	27	43
prescriptive	conjunctive	87	91
prescriptive	disjunctive	37	41

when one compares the proportion of participants who selected the abnormal cause as a function of the causal structure (as shown in Figure 4.5) to the proportion of participants who had a preference for the abnormal event in their normality inference. To determine the latter, we simply calculated the proportion of participants who exhibited a preference for the abnormal cause (i.e. whose judgment was greater than 0 in Figure 4.6). Table 4.3 shows that there is a very close correspondence between the percentage of participants who selected the abnormal cause as a function of the type of norm and the causal structure (selection), and the percentage of participants who inferred that a selected cause was abnormal (inference).

4.3.2.1 Individual Differences

The tight relationship between causal selections and normality inferences is also demonstrated by breaking down participants' normality inferences based on whether they themselves selected the abnormal or normal cause as a function of the causal structure (see Figure 4.7). Generally, participants tended to interpret an explanation in line with what they themselves would have said in the same situation. Interestingly, participants who themselves selected the abnormal cause in the conjunctive condition, had a stronger preference to infer the abnormal event than participants who selected the abnormal cause in the disjunctive condition. Moreover, as already apparent from Figure 4.6, there is an asymmetry in participants' inferences. Participants who selected the abnormal cause in the conjunctive structure (left-most data points in Figure 4.7) are more certain in their inference compared to participants who selected the normal cause for the disjunctive structure (right-most data points).

4.3.3 Discussion

The results of Experiment 1 are in line with previous literature on causal judgments (Gerstenberg & Icard, 2019; Icard et al., 2017; Kominsky et al., 2015), showing that the selection of causal explanations is affected both by normality and causal structure. People tend to select an explanation citing an abnormal cause in a conjunctive causal structure, but are more likely to select a normal cause when the structure is

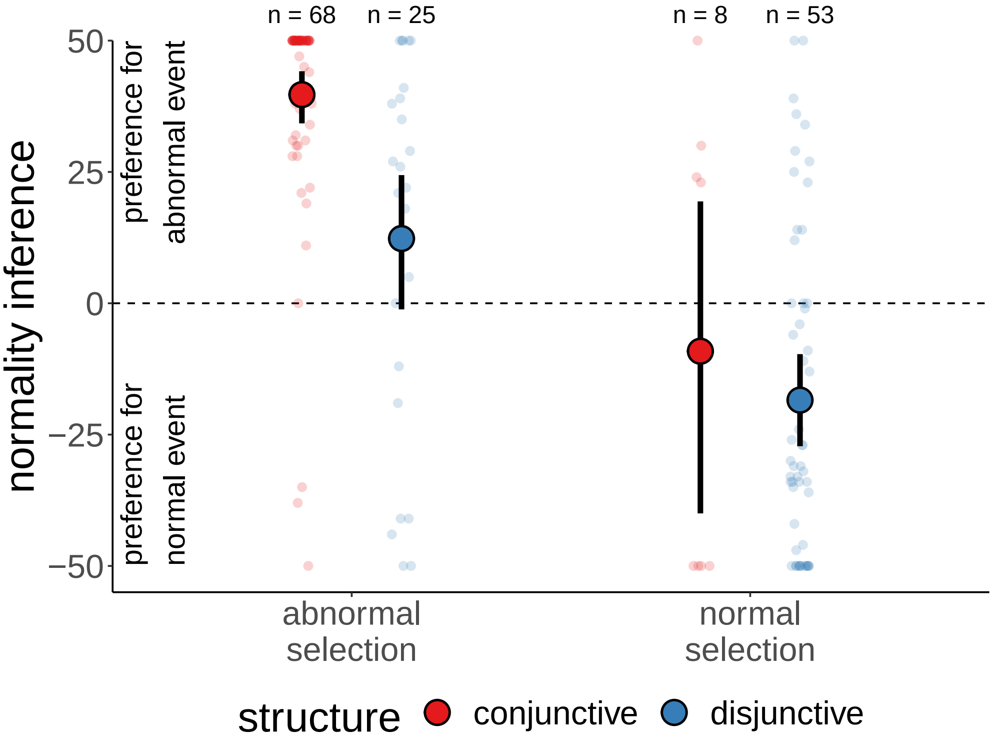


Figure 4.7: Experiment 1: Participants' preference for the abnormal cause (top) versus normal cause (bottom) as a function of the causal structure (conjunctive vs. disjunctive) and whether they themselves selected the abnormal or normal cause (abnormal vs. normal selection). For example, the left-most data points show participants' inferences who themselves selected the abnormal cause in the conjunctive scenario. *Note*: Large circles are group means. Error bars are bootstrapped 95% confidence intervals. Small circles are individual participants' judgments (jittered along the x-axis for visibility).

disjunctive ("Hypothesis 1"). Crucially, the experiment also confirmed our prediction of people's normality inferences from explanations ("Hypothesis 2"). People are more likely to infer that the cited event in the explanation was abnormal when the underlying causal structure is conjunctive, compared to disjunctive.

In general, people's normality inference closely mirrored their own explanation preferences. People were more likely to infer that a cause was abnormal if they themselves selected an explanation citing an abnormal cause before. Interestingly, people's prior causal explanation not only influenced whether they inferred an abnormal or normal cause, but also the strength of their inference. The asymmetry in participants' causal selections reported previously (Icard et al., 2017; Kominsky & Phillips, 2019) also shows up in their inferences. In line with what has been found previously (cf. Gerstenberg & Icard, 2019), participants' tendency to select the abnormal cause in conjunctive structures is stronger than their preference to select the normal cause for disjunctive structures (cf. Figure 4.5). Correspondingly, participants were more certain that the cited cause in the explanation was abnormal in the conjunctive scenario, compared to how certain they were that the cited cause was normal in the disjunctive scenario.

To conclude, Experiment 1 not only showed that normality and structure affect causal explanations, we also found that explanation and structure guide people's inferences about normality. In Experiment 2, we test whether participants can infer the causal structure of a scenario based on whether a normal or abnormal event was cited in the explanation.

4.4 Experiment 2: Inferring Causal Structure given Normality

In Experiment 2, we test whether participants can infer the causal structure of a scenario based on whether a normal or abnormal event was cited in the explanation.

4.4.1 Methods

4.4.1.1 Participants and Design

We recruited 213 participants (Mean$_{\text{age}}$ = 34, SD$_{\text{age}}$ = 10, N_{female} = 70, $N_{\text{undisclosed}}$ = 1) via Amazon Mechanical Turk (Crump et al., 2013). 70 participants were excluded for failing one or more exclusion criteria specified below, leaving a final sample of 143 (32.9% excluded). The experiment has a 2 explanation normality (normal vs. abnormal) × 2 norm type (statistical vs. prescriptive) design. Norm type and the normality of the cited cause in the inference task were manipulated between participants. The participants were randomly assigned to one of the four experimental conditions, *statistical normality & abnormal explanation* ($N = 33$), *statistical normality & normal explanation* ($N = 27$), *prescriptive normality & abnormal explanation* ($N = 41$), and *prescriptive normality & normal explanation* ($N = 42$). In this experiment, participants were instructed about both causal structures, and the selection task was presented for both causal structures as well.

4.4.1.2 Statistical Normality: Selection Task

The introduction to the statistical normality condition in Experiment 2 was largely the same as in Experiment 1. Participants received a text and diagram instruction about the billiard ball setup (Figure 4.2). However, rather than being introduced to only one of the conjunctive or the disjunctive billiard ball structure, participants learned about both structures. In contrast to Experiment 1, we didn't vary the position of the two blockers. In both the conjunctive and disjunctive setup, the dark red and light red blocker were always at the same position. Which blocker was on the top and which was at the bottom was randomised across participants.

Participants then proceeded to watch two video clips in which both balls went through the blockers and ball E went through the gate. One video clip showed the scenario in a conjunctive setup, and the other the disjunctive setup. As in Experiment 1, participants first had to make prediction judgments when the clip was paused shortly after the beginning, and then select a causal explanation after the full clip finished playing. Participants watched a clip that showed both the likely and unlikely ball hitting ball E in a conjunctive (or disjunctive) structure. Based on this clip, they then had to select either an explanation referring to the abnormal event, or an explanation referring to the normal event. Subsequently, participants were asked to do the same task again – this time with the alternative causal structure. The order of the clips with the conjunctive and disjunctive causal structure was randomised.

4.4.1.3 Statistical Normality: Inference Task

In the final inference task, participants received a diagram of a billiard ball scene in which both ball A and ball B went through the blocker and ball E went through the gate. However, the largest part of the billiard scene was grayed out (see Figure 4.3c). The causal diagram was missing the crucial information about where ball E was positioned at the beginning. Hence, the scene did not give away whether the causal structure was conjunctive or disjunctive.

Participants were told that Ben, a fictional participant, has witnessed the entire scene and selected the explanation “Ball E went through the gate because ball A [B] went through the blocker.". We counterbalanced across participants whether Ben's explanation referred to the abnormal or normal cause. Participants were presented with the following question: "Given Ben's decision, which of these two scenes did he see?". One endpoint of the slider showed a billiard scene with a conjunctive setup and the other endpoint a disjunctive setup. The endpoints of the slider were labeled "Definitely this one" and the midpoint "Uncertain". The left/right position of the scenes was randomised across participants.

4.4.1.4 Prescriptive Normality: Selection Task

Similar to Experiment 1, participants were instructed about the two employees Billy and Suzy. This time, however, they were informed that there are two offices in which Billy and Suzy sometimes work, depending on availability. The "Two-Door-Office" has a motion detector with a conjunctive structure, and the "One-Door-Office" has a motion detector with a disjunctive structure (see Figure 4.4). Given the confidentiality of the project, their boss sometimes instructs one of them to come into the office at 9am in the morning, while the other one is not allowed to come in that morning. In contrast to Experiment 1, normality was fixed: Who was allowed to come in and who not was *always* the same, independent of which office Suzy and Billy were currently working in. It was randomised across participants who of the two employees was supposed to come in, and who was supposed to stay away.

Participants then watched two video clips about two subsequent days in the company in which Billy and Suzy were given their instructions and both came in the next morning. One clip showed the scenario in the "Two-Door-Office" (conjunctive), and the other in the "One-Door-Office" (disjunctive). As in Experiment 1, participants made prediction judgments first (assessing their comprehension of the norms and causal structure), and then select a causal explanation. The order of the two clips was randomised.

4.4.1.5 Prescriptive Normality: Inference Task

Participants received a diagram showing a scene in which both Billy and Suzy came into the office at 9am in the morning and the motion detector went off. However, the entire floor of the office including the furnishing and door front was left blank. As a result, the diagram did not show whether they entered the “Two-Door-Office” office with the conjunctive motion detector or the “One-Door-Office” with the disjunctive motion detector. As in Experiment 1, fictional participant Ben had witnessed the scene and selected the causal explanation “The motion detector went off because Billy [Suzy] entered the office.” We counterbalanced across participants whether Ben’s explanation referred to the abnormal or normal cause. Participants were asked to indicate which scene they thought Ben had witnessed. A slider showed the scene in the two possible offices, together with the boss’ instructions for that day, on each endpoint respectively. The endpoints of the slider were labeled “Definitely this one” and the midpoint “Uncertain”. Which office scene was depicted left or right was randomised.

4.4.2 Results

Figure 4.5b shows participants’ selections as a function of the causal structure and norm. Note that this time, we manipulated the causal structure within participants, so we asked each participant to indicate their selection for both causal structures. Table 4.4 shows the pattern of selections. Most participants ($n = 80$) selected the normal cause when the causal structure was disjunctive, and the abnormal cause when the structure was conjunctive. There was also a large group of participants ($n = 44$) who selected the abnormal cause for both structures.

ANOVA results show that there was a significant effect of structure on causal selections $\chi^2(1) = 26.44, p < .001, \phi_c = 0.30$, 95% CI $[0.19, 0.42]$, as well as

Table 4.4: Experiment 2 – Causal selection patterns: Number of participants (n) for each possible combination of selecting the normal (or abnormal) cause for disjunctive and conjunctive structures. For example, there were 80 participants who selected the normal cause in the disjunctive causal structure and the abnormal cause in the conjunctive structure.

disjunctive	conjunctive	n
abnormal	abnormal	44
abnormal	normal	3
normal	abnormal	80
normal	normal	16

a significant interaction between structure and norm $\chi^2(1) = 4.19, p = .04, \phi_c = 0.12$, 95% CI $[0.00, 0.24]$. Overall, participants were more likely to select the abnormal cause for conjunctive causal structures (87%) compared to disjunctive structures (33%; see Table 4.5). The difference in participants' selections between the conjunctive and disjunctive structures was stronger in the prescriptive norm condition compared to the statistical norm condition (see Figure 4.5b).

Figure 4.8 shows participants' inferences about the causal structure of the situation as a function of the type of explanation (citing an abnormal or a normal event) and the type of norm (statistical or prescriptive). Participants' inferences were affected by the normality of the explanation (see Table 4.6), $F(1, 139) = 77.15, p < .001, \eta_p^2 = 0.36$, 95% CI $[0.24, 0.47]$ Participants had a stronger preference to infer the conjunctive structure for explanations citing an abnormal event $(\mathrm{M} = 74.07, \mathrm{SD} = 30.26)$ compared to a normal one $(\mathrm{M} = 27.25, \mathrm{SD} = 31.96)$.

Table 4.5: Experiment 2 – Causal selection: Analysis of variance of the causal selection results.

model specification: `selection ~ structure * norm + (1 | participant)`

Term	Chisq	Df	Pr(>Chisq)	Cramers_v	CI	CI_low	CI_high
structure	26.44	1.00	0.00	0.30	0.95	0.19	0.42
norm	0.17	1.00	0.68	0.02	0.95	0.00	0.14
structure:norm	4.19	1.00	0.04	0.12	0.95	0.00	0.24

Table 4.6: Experiment 2 – Structure inference: Analysis of variance of the normality inference results.

model specification: `structure rating ~ 1 + explanation * norm`

Term	Sum Sq	Df	F value	Pr($> F$)
explanation	76137.49	1.00	77.15	0.00
norm	34.48	1.00	0.03	0.85
explanation:norm	21.84	1.00	0.02	0.88
Residuals	137174.06	139.00		

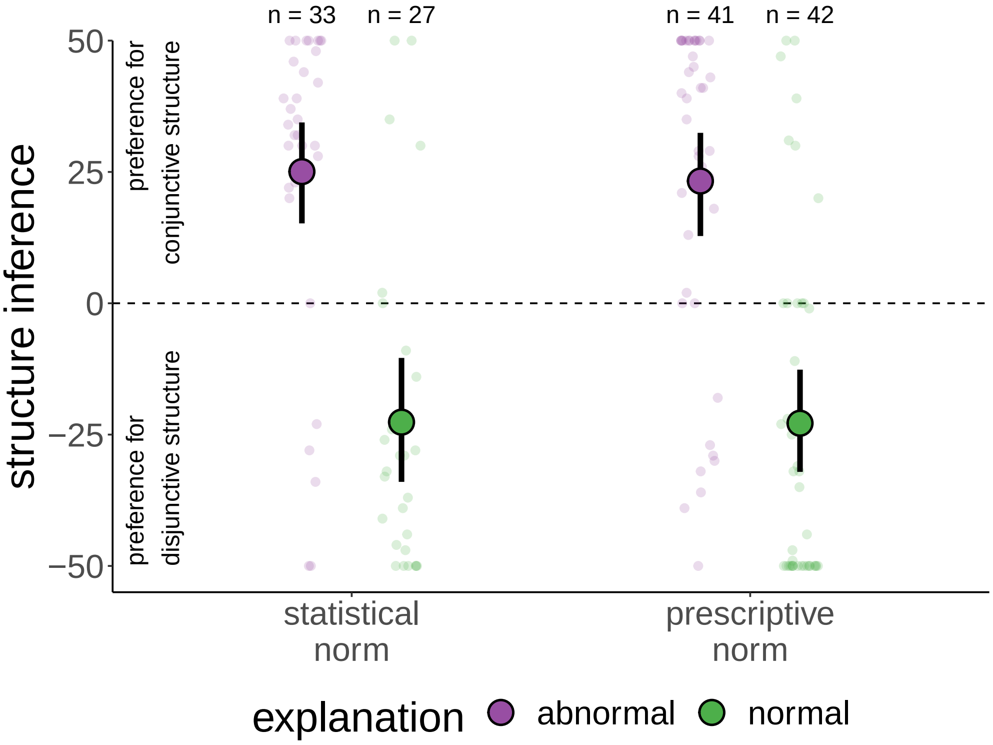


Figure 4.8: **Experiment 2**: Participants' preference for the conjunctive (top) versus disjunctive (bottom) structure as a function of the explanation (abnormal cause vs. normal cause) and the norm type (statistical vs. prescriptive). *Note*: Large circles are group means. Error bars are bootstrapped 95% confidence intervals. Small circles are individual participants' judgments (jittered along the x-axis for visibility).

4.4.2.1 Individual Differences

Figure 4.9 shows participants' structure inferences depending on what causal selections they themselves made. For example, the left-most data show the inference that participants made based on an abnormal explanation who themselves selected the abnormal cause both for the disjunctive (D) and conjunctive structure (C). What stands out is that the strength of the inference is strongest for the largest group of participants ($n = 80$) who selected the normal cause for the disjunctive structure and the abnormal cause for the conjunctive structure. For this group of participants, the difference between the group's mean inference based on an abnormal versus normal explanation is largest (Mean = 59.03). Even those participants who selected the abnormal event for both structures, or those who always selected the normal event, still had a preference for the conjunctive structure for abnormal explanations,

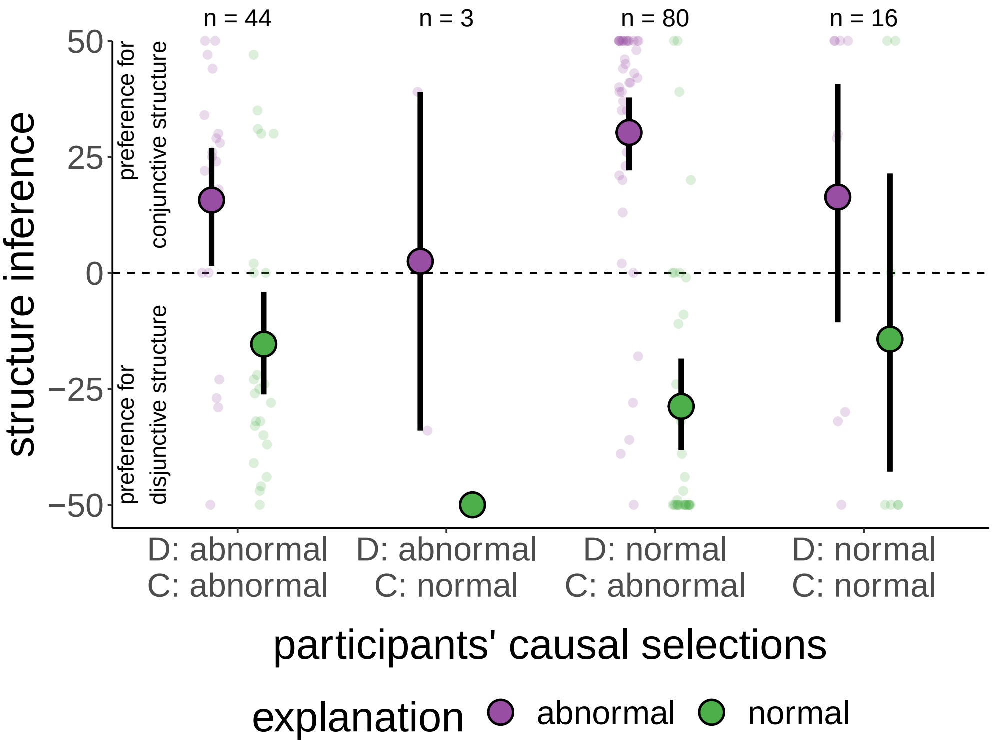


Figure 4.9: Experiment 2: Participants' preference for the conjunctive (top) versus disjunctive (bottom) structure as a function of the explanation (purple = abnormal cause, green = normal cause) and what causal selections they themselves made (D = disjunctive, C = conjunctive). For example, the left-most data points show participants who selected the abnormal cause both for disjunctive and conjunctive structures, and who made a structure inference based on an explanation citing an abnormal cause. *Note*: Large circles are group means. Error bars are bootstrapped 95% confidence intervals. Small circles are individual participants' judgments (jittered along the x-axis for visibility).

and the disjunctive structure for normal explanations. Here, however, the difference between the inferences as a function of whether the explanation was abnormal or normal was weaker (M = 31.04 for the group of participants who always selected the abnormal cause, and M = 30.62 for the participants who always selected the normal cause).

4.4.3 Discussion

Experiment 2 replicated the causal explanation findings from Experiment 1, but this time in a within-participant design. Overall, most participants chose the explanation referring to the abnormal cause in a conjunctive causal structure, but referred to the normal cause in a disjunctive structure. Experiment 2 also confirmed our predictions about people's inferences from a causal explanation when the normality of the cause is known (Hypothesis 3). People were more likely to infer a conjunctive causal structure, rather than a disjunctive structure, when the cited cause was abnormal.

As predicted, how certain participants were about their inference depended on their own causal explanations (Hypothesis 4). Virtually all participants inferred a conjunctive structure when an explanation referred to an abnormal cause, and a disjunctive structure when a normal cause was cited. This inference was strongest for participants who themselves selected the abnormal cause in the conjunctive structure, and the normal cause in the disjunctive structure. In contrast, for participants who deviated from this pattern, the structure inference was weaker.

The results of Experiment 2 show that the influence of norms and causal structure on causal explanations persists when people are able to directly compare and select explanations for both types of causal structures. People generally infer the causal structure from the normality of a cause in a way that tracks the interaction between normality and causal structure on causal explanations found in the literature (Gerstenberg & Icard, 2019; Icard et al., 2017; Kominsky et al., 2015). However, whether people themselves would give these explanations has an impact on the strength of their structure inferences. Participants who themselves selected causal explanations that matched the inference pattern (Abnormal: C, Normal: D) showed the strongest inference compared to those who selected deviating causal explanations. These results in particular shed new light on the crucial role of explanatory preferences for inferences.

4.5 General Discussion

As Hilton (1990) observes, "the verb 'to explain' is a three-place predicate: *Someone* explains *something* to *someone*." (p. 65). Indeed, the communicative dimension of explanation is essential. In this chapter, we have taken a first step in investigating the inferences that people make when facing explanations in communicative settings.

As a case study, we focused on the role of normality and causal structure in explanations. In line with previous literature, we show that people prefer to explain an outcome in a conjunctive causal structure by referring to an abnormal cause. In contrast, in disjunctive causal structures, people prefer to cite a normal cause in their explanation. Crucially, we show that these two factors not only influence what explanations people give, they also determine the kind of inferences people draw from the explanations of others. When provided with a causal explanation about what happened and information about the causal structure, people are able to infer the normality of the cited cause. Likewise, people infer the causal structure of a scenario when provided with an explanation together with background information about the normality of the cited cause. We show this pattern for both statistical as well as prescriptive normality. These results are consistent with the idea that a listener considers what they themselves would have said in the given situation, and interprets a speaker's explanation accordingly.

In the remainder of this chapter, we will discuss how our findings relate to prior work on explanation. In particular, we aim to revisit current theoretical accounts in the debate around norms in causal cognition, however, with a different focus. In particular, we want to examine how well each of these accounts fit with the idea that the judgments and explanations that are influenced by norms have this communicative dimension that is so crucial for the inferences that we draw from them. Finally, we want to discuss the broader role of theory of mind for inferences from explanation.

4.5.1 Explanations in Communication

The communicative role of explanation in human affairs is well recognised in the existing literature, and researchers have shown a number of notable patterns in the ways that a typical speaker will adapt to their audience. For instance, what begs an explanation in the first place is often a "why"-question, either implicit or explicit. The crucial ability of theory of mind enables such effective communicative transfer. An adequate explainer bridges the difference between what happened and what kind of background expectations or knowledge the recipient of the explanation holds (Bruckmüller, Hegarty, Teigen, Böhm, & Luminet, 2017). People flexibly adapt an explanation when their counterpart is ignorant or shares only a partly overlapping view of the explained event (Slugoski et al., 1993). People also omit a factor in an explanation if they consider it irrelevant or redundant to the question they have been asked (Einhorn & Hogarth, 1986; Hilton & Jaspars, 1987).

Our focus here has been on the role of norms and causal structure. One critical ingredient that allows people to communicate so much so succinctly (merely uttering "*E* because *C*") is that we largely share intuitions about the interactions of norms and causal structure. For instance, people seem to share the judgment that it is more appropriate to mention a normal cause in a disjunctive setup than in a conjunctive setup. Through quite general communicative principles and the reference to the mental states and preferences of others, such shared intuitions facilitate strikingly efficient information transfer. Simply by thinking about what the other person would deem an appropriate utterance in different possible states, a listener is able to infer the actual state. This suggestion raises two key questions:

1. Why do people largely share these intuitions in the first place?
2. Why do norms and causal structure affect causal explanations in the way that they do?

It is tempting to speculate that the answer to 1. is precisely that shared causal intuitions allow the type of communicative efficiency that we demonstrated in this chapter. However, as far as communicative coordination is concerned, there is evidently nothing that singles out this specific pattern as especially efficacious. Indeed, if all preferences were flipped (e.g., preferring normal causes in conjunctive rather than disjunctive situations), people would be able to make all the same inferences, on the present account. Thus, unless we believe that the specific patterns are truly arbitrary and, for example, arose purely by chance, this answer to question 1. does not yet offer a fully satisfying answer to question 2.

As discussed throughout this thesis, there are already a number of proposals about 2. in the existing literature. It is thus worth considering how the communicative dimension of explanation studied here might interact with prominent accounts of the role of norms in causal judgment. We will consider again two accounts that foreground different assumptions about what mediates the effect of norms on causal explanations, but also add a new perspective here: 1) counterfactual reasoning, 2) blame and accountability, and 3) optimal interventions.

4.5.2 Counterfactual Reasoning

The *counterfactual reasoning* account (Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Knobe, 2009; Kominsky et al., 2015) draws on a substantial psychological link between causal explanation and *counterfactual relevance* (Byrne, 2016; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Kominsky & Phillips, 2019; Phillips et al., 2015). Some counterfactual possibilities strike us as more relevant than others, and perhaps also come to mind more readily (see also mental model theory P. Johnson-Laird & Khemlani, 2017). Specifically, abnormal events tend to trigger counterfactual thoughts about what would have happened had things gone normally, while the reverse does not seem to hold (Kahneman & Miller, 1986). The relative availability of normal alternatives for abnormal causes makes these counterfactual necessity claims more easily verifiable, which in turn strengthens the relevant causal claim.

On some formalisations of the counterfactual reasoning account, a causal claim "E because C" incorporates not only necessity, but also notions of sufficiency or stability, for example, the extent to which E would still have resulted from C had background conditions been slightly different (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015b; Grinfeld et al., 2020a; Icard et al., 2017; Kominsky et al., 2015; Pearl, 1999; Vasilyeva, Blanchard, & Lombrozo, 2018; Woodward, 2006). (Icard et al., 2017) specifically predict the most prominent patterns studied in the existing literature, including those investigated in the present work.

If the counterfactual reasoning account is correct, then norm effects and the interaction between norms and causal structure arise simply from the way our minds work. Given which counterfactuals are more easily available, certain causes come more easy to mind and are more readily cited than others, establishing the known patterns in causal explanations and judgments over time. It is this pattern, the product of our cognitive machinery, that shapes our communication about causes over time. Thus, one possibility consistent with this type of account is that communicative efficiency is just a serendipitous byproduct of a more basic psychological pattern.

4.5.3 Blame and Accountability

A different line of research that has already been extensively discussed in previous chapters aims to explain the role of normality in causal selections in terms of blame and responsibility attributions. Some have argued that people's causal judgments are biased by a desire to assign blame to the abnormal factor (Alicke, 2000). According to this account, emphasising the causal contribution of an abnormal cause allows people to validate their spontaneous blame response. Others argue that people's ordinary concept of causation is itself normative, with causal judgments being akin to judgments about responsibility (Sytsma, 2019a; Sytsma & Livengood, 2019; Sytsma et al., 2012). Samland and Waldmann (2016) contend that these effects arise due to pragmatic factors in the context of norm violations and human agents. Rather than assessing an "actual causal" process, participants interpret the causal test question as a request to assign accountability (Samland & Waldmann, 2014, 2015, 2016). Accordingly, a speaker uses a causal explanation "*E* because *C*" to communicate some form of attribution of responsibility or blame. On this type of account, we should therefore expect participants to make inferences consistent with the cited cause being blameworthy in some way.

While blame-oriented accounts explicitly address the communicative function of causal explanations, and provide a plausible explanation for causal selections in case of prescriptive norms and human agent causation, as discussed in Chapter 2, it is less clear how they would work for inanimate objects and statistical normality. When provided with the explanation "Ball E went through the gate because Ball A went through the blocker", it seems questionable whether the recipient will interpret this statement as an expression of blame or responsibility attribution centered on the ball. Henne, Kulesza, et al. (2021) demonstrate the effect of norms on people's causal judgments about billiard balls even after controlling for perceived agency. Moreover, the effect of normality has been shown for outcomes that are positive, neutral and bad in nature (Icard et al., 2017; Reuter et al., 2014). In addition, these accounts leave open why in a disjunctive structure, people blame the agent that adheres to the prescriptive norm. However, at the current theoretical state, these blame-oriented models do not explain how a normative judgment is made in circumstances other than a clear rule violation, or when an action results in a bad outcome (Alicke & Rose, 2012a; Samland & Waldmann, 2016). Without a more precise account of responsibility or blame, it doesn't seem possible to identify a sensible and unequivocal blame response from the diverse range of causal explanations that are impacted by normality. This makes it difficult for accounts referring to blame or accountability to predict the overall pattern found in our experiments, even the basic patterns for the selection tasks.

4.5.4 Explanations point out Optimal Interventions

The previous accounts have focused on "backward-looking" aspects, such as how our explanatory practices relate to assessments of responsibility and blame (Lagnado et al., 2013; Malle, Guglielmo, & Monroe, 2014b; Woodward, 2011). As discussed in Chapter 3, it has recently been suggested in philosophy and psychology that causal judgments and explanations additionally have an important "forward-looking" function: a good explanation helps us pinpoint useful places for future intervention and action (Chi, De Leeuw, Chiu, & LaVancher, 1994; Gerstenberg & Icard, 2019; Hitchcock, 2012; Liquin & Lombrozo, 2020; Lombrozo & Carey, 2006; Woodward, 2003). Simply put, good explanations should not just be convincing, they should also lead to positive downstream effects (Danks, 2013; Woodward, 2014). On this view, causal explanations are used to identify *optimal points of intervention* (Hitchcock, 2012; Hitchcock & Knobe, 2009; Lombrozo, 2010; Morris et al., 2018; Woodward, 2006). A speaker is assumed to highlight for a listener some variable that is especially worthy of attention for the purpose of future decision making. An optimal point of intervention may certainly be a variable that is in some way deserving of blame or censure, and in this sense the account is consistent with a blame account. At the same time, the optimal intervention account need not be tied to blame or accountability per se. Rather, what makes for a good or promising point of intervention may vary from context to context, and importantly, such considerations will often be pertinent even when assessment of blame is inappropriate.

How might an optimal intervention account shed light on the results in this chapter? Specifically, in what sense might it be better to intervene on an abnormal cause in conjunctive structures, and a normal cause in disjunctive structures? For this, it matters how people construe the notion of an optimal intervention. Consider our billiard ball setting with a conjunctive structure: Ball A and ball B both need to pass their blockers in order for ball E to go through the gate. Suppose ball A has a 20% chance of going through the blocker, while ball B has an 80% chance. For example, it's possible that people think an optimal intervention is the one that is most likely to make the outcome happen (optimal intervention $= max(p(E|do(C)))$, that is, to make the ball go through the gate. Here E stands for the event of ball E going through the gate. The do() operator indicates that we fix the event via an intervention (making it either true or false). In that case, for conjunctive structures, one should intervene on the abnormal event, ball A and make it less "abnormal", i.e. increase the likelihood of ball A to go through the blocker. In contrast, for disjunctive structures it doesn't matter since either event is sufficient to make the outcome happen.

Alternatively, suppose that an optimal intervention is one that makes the largest difference to the probability of the outcome of interest. We want to compare $p(E|do(C)) - p(E|do(\neg C))$, where C is either ball A or ball B going through the block. There is an 80% chance that ball E will go through the gate if we make sure that ball A goes through the blocker, and a 0% chance if we prevent ball A from going through the blocker. In contrast, for ball B we get $p(E|\text{do}(B)) - p(E|\text{do}(\neg B)) = 0.2 - 0 = 0.2$. Thus, in a conjunctive scenario intervening on the less likely event makes the biggest difference to the probability of the outcome. To make the biggest difference to the outcome, a person should intervene on ball A rather than on ball

B.[3]

The communicative framework sketched in the introduction focused on the resolution of uncertainty about a situation, and our experiments similarly highlight this epistemic aspect of linguistic interpretation. However, it may be that the ultimate explanation for the specific patterns of norm effects we see on causal judgments is properly framed in a broader communicative story. The optimal intervention framework also offers another potential insight into the asymmetric way in which norms and causal structure affect people's causal selections. Recall that participants are more likely to select the abnormal cause in conjunctive structures than they are to select the normal cause in disjunctive structures (see Figure 4.5), and that this asymmetry in how explanations are generated is reflected in the inferences that people draw (see Figure 4.6). How may this pattern of causal selections arise from communicative pressures?

Suppose that causal judgments are sensitive to other communicative pressures aside from those discussed above. In particular, it seems reasonable to assume that identifying an abnormal event will often be helpful, especially when the listener is unaware of it. After all, when an alternative, *normal* event can reasonably be assumed, mentioning the abnormal event will be strictly more informative. We thus have (at least) two communicative pressures that may shape causal explanations: being generally informative and highlighting a variable that would be a good point of intervention. In conjunctive structures these two pressures both focus attention on the abnormal event. However, in disjunctive structures they pull in different directions. These conflicting pressures may account for the asymmetric pattern observed in the data.

Of course, as has been long appreciated in the literature (see, e.g., Coffa, 1974), a purely forward-looking approach to causal explanation, focusing only on a variable's future causal potential, may flounder on cases where backward-looking and forward-looking dimensions come apart. If a person takes a treatment for an illness

[3] By the same logic, it's better to intervene on the *more likely* cause in a disjunctive scenario. Here, for A we get $p(E|\text{do}(A)) - p(E|\text{do}(\neg A)) = 1 - 0.8 = 0.2$, and for B we get $p(E|\text{do}(B)) - p(E|\text{do}(\neg B)) = 1 - 0.2 = 0.8$. Thus, in the disjunctive scenario intervening on the more likely event, B, makes the bigger difference to the probability of the outcome.

but then gets better independently, the treatment should not be part of the explanation, no matter how effective it is in general. What matters is not how good of an intervention we expect taking the treatment to be in the future, but rather the fact that *in this particular instance* the treatment did not cause the improvement. Negotiating the balance between these two dimensions of causal explanation is an important challenge for any adequate theory of explanation (cf. Hitchcock, 2017).

4.5.5 Theory of Mind and Inferences from Causal Explanations

As argued before, the reference to the speaker's mental states bolsters the listener's ability to infer systematic information from an utterance that goes beyond what is explicitly said. In turn, if the speaker assumes that the listener has some knowledge about the causal scenario, e.g. about the causal structure or the normality of the causes, they can expect the listener to infer the missing piece of information from an explanation. However, in realistic scenarios we cannot always expect that listeners (or speakers for that matter) have full knowledge of the situation. For example, suppose that our listener knows neither the causal structure nor the normative status of variables. From an utterance "E went through the gate because A went through its blocker," such a person could at best infer that, either the causal structure is conjunctive and A was unlikely, or the structure is disjunctive and A is likely. Likewise, the listener would cease to make any systematic inferences if they knew the speaker to actually have partial or no knowledge about the causal scenario. If Ben did know that a conjunctive motion detector structure was in place, but did not know whether Suzy or Billy was supposed to come in that day, his explanation "The motion detector went off because of Billy" will not give rise to any systematic inferences about the normality status of the agents. Crucially, whether and to what extent inferences can be drawn are highly dependent on both speaker's as well as listener's epistemic states.

Our main proposal here draws on the assumption that people simulate what would have been most reasonable or sensible for a speaker to say, using the pervasive influence of norms on causal judgments as a test case. In absence of knowledge about the exact explanatory preferences of the speaker, these are filled by what the listener themselves would have deemed appropriate. These explanatory preferences however can go well beyond the influence of normality, and a listener might take into account what they know about speaker's general knowledge, beliefs, moral views, etc. Take for example the statement "Brexit happened because of one conservative politician!'. Depending on whether you take the speaker to be a pro-Brexiteer, or an Anti-Brexit campaigner, you might draw different inferences about who is meant here. In case of assumed pro Brexit attitudes, you might infer that Boris Johnson is meant, the conservative politician who campaigned in favour of an exit from the European Union. In contrast, if you take the speaker to be a pro-European "Remainer", you might infer the named causal agent to be David Cameron, who wanted to remain in the EU, but failed to run a successful campaign convincing the British people of the advantages of staying.

Even the inference about abnormal causes might be susceptible to assumptions about the speakers views about morality and normality. Knowing whether a person endorses consequentialist, or rather deontological ethics might lead to different inferences from their utterance "Switching the tracks was morally wrong!" about the number of saved vs. killed people in a trolley scenario. What is judged as "abnormal" might be highly dependent on what the speaker takes to be morally prescriptive, and a causal explanation such as "This party is the cause of the current crisis in the US" might lead to very different inferences depending on the speaker's conservative vs. liberal views. The same holds for our understanding about descriptive normality. Imagine that, during a rainy Californian summer, the regular irrigation together with heavy rainfall has caused Stanford University's green areas to swamp. When talking to a gardener from Europe who pictures California as the all-year sunshine state, we might refer to the unusual rain fall this summer as the cause for the flooding. In contrast, when talking to a local who has witnessed the heavy rainfall this summer, the regular irrigation might be more appropriate to explain the incident. The fact that both interlocutors take into account each other's mental state for an efficient information transfer hence underlines the crucial role of theory of mind in this endeavour (P. Grice, 1989).'

4.5.6 Future Directions

In this chapter, we focused on people's inferences about event normality and causal structure. Future work will examine to what extent our communication-theoretic account can capture a broader array of inference patterns. Hence, any factor that systematically influences what kind of causes people select and deem appropriate to report in an explanation is a worthwhile object of investigation for inferences from explanations. For example, research has shown that people often prefer to cite early or late events as the cause of the outcome in a way that is sensitive to causal structure (Brickman, Ryan, & Wortman, 1975; Gerstenberg & Lagnado, 2012; Glautier, 2008; Hilton, McClure, & Sutton, 2010; McClure et al., 2007; Spellman, 1997). Given this systematic pattern, we would predict that people should be able to infer the temporal order of events from explanations given knowledge about the causal structure. As another illustrative case study, we are interested in exploring the kinds of inferences that people draw from social evaluations. In the social domain concepts like *blame* are closely related to explanation (see, e.g. Gerstenberg et al., 2018b; Malle, 2021). Again, if judgments of blame serve partly a communicative function (letting someone know that one expects for their behaviour to change in the future), we would expect that people can make inferences about what happened from such social evaluations (Davis, Allen, & Gerstenberg, 2021). Finally, a large part of this thesis has made the case for the impact of mental states on making causal judgments and explanations. If all the causal aspects in a scenario are known, a causal explanation such as "Billy caused the motion-detector to go off" might still lead us to infer something about Billy's mental states, for example, his knowledge of or desire towards the outcome. Causal explanations referring to causal agents might allow to us infer about the causal agent's epistemic states. This, too, remains to be investigated.

4.6 Conclusion

In this chapter, we have investigated the communicative and epistemic dimensions of explanation, revealing some of the rich and subtle inferences people draw from them. We find that people are able to infer additional information from a causal explanation beyond what was explicitly communicated, such as causal structure and normality of the causes. Our studies show that people make these inferences in part by appeal to what they themselves would judge reasonable to say across different possible scenarios, drawing on the fundamental skill of theory of mind. The overall pattern of judgments and inferences brings us closer to a full understanding of how causal explanations function in human discourse and behaviour, while also raising new questions concerning the prominent role of norms in causal judgment and the function of causal explanation more broadly.

Chapter 5

General Conclusion

“Rerum cognoscere causas” – “to know the causes of things” – is the motto of various universities and engraved over lecture halls, school, or library buildings. As abbreviation of the famous verse of Latin poet Virgil (“Fortunate, who was able to know the causes of things”), it captures the two central components that represent the major topics in this thesis: causality and mental states, more specifically, our ability for theory of mind. On the one hand, we strive to learn and know the causal laws and relationships in the world. To know the causes of things imparts power: We can explain and understand the world around us and effectively engage with it and change it for the better. However, when we apply our causal knowledge in the social world, we do so in combination with our theory of mind. We attribute causality based on what agents know and could have known, and based on the mental states we infer from their behaviour. We also adapt our causal explanations to what the person knows, and infer the kind of information we intuit a speaker would find most appropriate to explain. We have argued that in the social world, “to know the causes of things” entails so much more than just what a causal agent did. In our view, our understanding of causality cannot be separated from our understanding of other people’s minds.

Before we begin to situate our findings into a broader context, we will briefly summarise the results of our three research projects again.

5.1 Summary

5.1.1 Chapter 2: Abnormal Agents and Epistemic States

In chapter 2, we took a closer look at the 'abnormal selection' problem (Hesslow, 1988) in agent causation. More precisely, we explored why it is that in conjunctive causation, people judge an atypically acting agent as more causal than an agent who acts as usual. We speculated that the typicality of agent behaviour, i.e. how frequently agents act, might also have an impact on their expectations about the consequences of their actions. In particular, in a causal structure that requires two actions of a certain type for the outcome, if one agent frequently performs the relevant causal action, the outcome occurs as soon as the other agent does the same. An agent who usually does not perform the respective action will hence be in a good position to foresee that the outcome will happen once they decide to behave 'abnormally'. Is it the epistemic asymmetry, rather then the asymmetry in normality, what drives the difference in people's causal judgments about an atypical and typical causal agent?

In four experiments, we investigated this question (Kirfel & Lagnado, 2021). In Experiment 1, we found that people judge an abnormal agent as more causal than a normal agent, but also attribute to the abnormal agent an epistemic advantage, that is, they attribute to the abnormal agent more knowledge about the other agent's behaviour. If this epistemic asymmetry between abnormal and normal agent is lifted (Experiment 2), people attribute equal causality to abnormal and normal agent. In Experiment 3, we demonstrate this finding for cases in which agents' knowledge about each other extends to novel contexts. Experiment 4 ultimately provided evidence for the last piece of our argument. The epistemic difference between abnormal and normal agent also has an impact on the their expectation about the outcome. The abnormal agent is judged to be able to foresee the causal outcome to a greater extent. Based on this epistemic asymmetry, given that the abnormal agent still acted, we show that people also infer outcome-oriented mental states to a greater extent for the abnormal vs. normal agent.

The results from Chapter 2 fit into a series of previous studies (Kominsky &

Phillips, 2019; Samland & Waldmann, 2016), suggesting that the preference for prescriptively abnormal causes, i.e. moral norm-violating agents, can only be interpreted in conjunction with the agents' epistemic states. We show that this also holds for statistical or *descriptive* norms, that is, the typicality or frequency of agent behaviour. While the influence of normality on causal judgments may not be entirely reduced to the epistemic states of abnormal agents, our experiments point to the crucial role of agent epistemic states in people's causal reasoning about agent causes. This raises the broader question of what role mental states occupy in social causal cognition. The experiments in chapter 3 aimed to provide an answer to that question.

5.1.2 Chapter 3: Epistemic Interventions in Causal Reasoning

Why is it that ignorant agents are not only judged as less blameworthy, but also less causal for the consequences of their actions? In chapter 3, we suggest that mental states like ignorance or knowledge have an inherent function in causal reasoning and judgment. Drawing on the long tradition of counterfactual dependence theories in causality, we suggested that under certain circumstances, people use counterfactuals over mental states, rather than causal actions. While counterfactual frameworks posit that people imagine what would happened if a causal agent had not acted, we hypothesised that people primarily test for the occurrence of the outcome in a scenario in which the agent has the relevant mental state, for example, knowledge about the outcome. We have argued that there are several theoretical advantages in assuming epistemic interventions in causal reasoning. On the one hand, epistemic states often function as natural preconditions for (the imagination of) changing actions (Walsh & Byrne, 2007b). At the same time, intervention on prior causes is associated with weaker counterfactual dependence and weaker causal strength, as reflected in the reduced causal attribution to ignorant agents. On the other hand, interventions on mental states represent effective interventions on adaptive agent behaviour and securing the desired causal outcome. Only an agent who knows about the positive consequences of their action, or who does not intend to bring about the bad outcome, might act accordingly and secure or avoid a certain course of events.

Epistemic interventions are a hence natural manifestation of the forward-looking function of causal judgments (Hitchcock, 2012; Lombrozo, 2010).

In four experiments, we found evidence for epistemic interventions in causal and counterfactual reasoning. We showed that epistemic states play a role in a variety of different epistemic contexts: Knowledge vs. ignorance (Experiment 1), self-caused vs. externally caused ignorance (Experiment 2), few vs. many epistemic actions (Experiment 3) and finally, in cases where an epistemic action did not result in knowledge, but could have under different circumstances (Experiment 4). While the idea of epistemic interventions is in principle compatible with a variety of theoretical frameworks, we think that it might have the potential to bridge the divide between more blame-oriented and counterfactual dependence theories of causation.

5.1.3 Chapter 4: Inferences from explanations

While years of research have been dedicated to studying the factors that influence judgments about causal structure and causal strength, very few studies have tested whether the reverse holds: that is, whether people are actually able to reverse-engineer those factors from causal judgements and explanations. Drawing yet again on normality as a crucial factor for causal judgments, this is what we did in the last research project of this thesis. The necessity to rely on the ability for theory of mind, i.e. understanding the beliefs, knowledge states or intentions of a speaker, in order to make inferences beyond what is explicitly stated, is becoming increasingly acknowledged in pragmatics (Bosco et al., 2018; Fernández, 2013; Zufferey, 2010). In two sets of experiments, we find that people are able to infer information about normality — both prescriptive and statistical — as well as the causal structure in a scenario from a causal explanation as short as ‘E because of C’. By combining theory of mind with some generic principles of pragmatics, we are able to offer an explanation as to *why* this is the case. People are able to infer additional information about the world by simulating different contexts and probing under which of those the explanation would have been the most appropriate utterance by the interlocutor. It is because we rely on others to communicate meaningfully with us that we can simulate what they themselves deem “the cause” in a causal scenario. This allows us to infer rich information from their relatively comprised utterances.

In the remainder of this thesis, we reflect on the implications of this work.

5.2 Theory of Mind and Causal Cognition

Our tendency to infer and take into account the mental states of others is crucial. We grow up in rich social environments, being socialised through the acquisition of a sophisticated theory of mind of the people around us. In this thesis, we have argued for a close connection between ToM and causal cognition (Lombard & Gärdenfors, 2021). We infer about the mental states from the normality of the agents' behaviour, and take them into account when making causal judgments. We fine-tune the attribution of causality in alignment with a causal agent's epistemic conditions or state of ignorance about the consequences of their actions. And we are able to infer a variety of non-causal information from the causal explanations people give to us by referring to their causal selection preferences. The relationship between morality and causality as well as their equivalent in human cognition has been discussed extensively in philosophy, psychology and the cognitive sciences. Across all three research projects, we have discussed how our findings inform the debate around the relationship of morality and causality in human cognition, often focused on the particular role of norms in causal judgments.

Ultimately, we think that the intersection of theory of mind and causal cognition that we have investigated in this thesis sheds a new light on this debate. What causal agents know, intend or believe forms an important starting point for various moral attributions about their behaviour: blame (Alicke, 2008), responsibility (Sytsma, 2019a), or accountability (Samland et al., 2016a). In that sense, mental states play a role for the backward-looking function of causal attributions and causal selection, that is, the basis for moral evaluations, the assignment of blame or credit, etc. Our causal judgements trace back from the outcome to the potential cause of an outcome, and further contextual information such as mental states or normative features inform our evaluative response. It has, however, been suggested in philosophy and psychology that causality also has an important "forward-looking" function:

causal selection pinpoints useful places for future intervention and action (Chi et al., 1994; Hitchcock, 2012; Liquin & Lombrozo, 2020; Lombrozo & Carey, 2006; Woodward, 2003). The factors that are singled out in causal judgments should also lead to positive downstream effects (Danks, 2013; Woodward, 2014). Mental states are the vehicle of agent behaviour and their potential consequences, and therefore present suitable targets for causal intervention. Even if interventions on mental states might represent more 'fuzzy' interventions than interventions on actions or omissions, they secure an enduring change of the outcome in the long run.

As such, mental states reflect both the backward-looking as well as the forward-looking aspect of causation, and might be able to bridge the two camps in the debate about the interpretative sovereignty of people's causal judgments. When taking into account mental states for causal judgments, people acknowledge both a more normatively laden as well as genuine causal-interventionist dimension of causality. In order to develop this debate further, future work will need to pit these two aspects of causation against each other, that is, causality as a path to blame vs. causality as a path to future intervention. Theory of mind sits at the intersection of causal and social cognition. As such, we envisage that the manipulation of agent mental states provides suitable candidate factors to achieve this.

5.2.1 Social and Causal Learning

Mental states are causally efficacious; we understand each other as acting because of desires, intentions, beliefs, etc. A large area in philosophy of mind is dedicated to the question of how it is that mental states can have causal properties, not *if* they do (Jackson, 1977, 1996; Kim, 2007; Yablo, 1992). In consequence, inferring other people's mental states also means an inference about non-visible or hidden causal variables of their behaviour. Making systematic inferences from other people's actions or action consequences requires an understanding that agents have mental states like knowledge, desires, and beliefs, but also that these mental states guide how agents act; and that these actions can somehow impact the environment. In this thesis, we have argued for a close connection of causal theory and theory of mind in people's judgments about causation. But are these two abilities also connected in the process of development and learning? Section 1.4 has already summarised important literature showing that children draw on ToM in causal learning. We will revisit a few selected studies again demonstrating that causal learning often leverages on social learning, and vice versa.

In a study by Goodman, Baker, and Tenenbaum (2009), participants were given scenarios in which a knowledgeable agent, a co-worker in a plant nursery, performs two simultaneous actions, e.g. pouring a blue and a yellow liquid over a set of flowers, and some time later, the flowers start to grow. To show that causal inferences are due to knowledgeable agent assumptions, Goodman et al. (2009) contrasted this case with a 'self condition' which described the actions as being taken by the participant ("One day when your coworker is gone, you find a yellow liquid and a blue liquid in his supplies and simultaneously pour them on the flowers."). Goodman et al. (2009) find that people infer a conjunctive structure, i.e. both actions necessary for the effect, to a greater extent in the knowledgeable agent vs. self condition. In contrast, if the agent only poured one liquid over the plants but drank the other, most people inferred a disjunctive structure. Drawing on the assumption that the causal agent both knew the causal structure and desired the outcome to happened, people are able to make systematic causal inferences about the underlying causal structure. Causal learning from goal-directed or — as investigated in chapter 4 — communicative actions differs significantly from causal inferences afforded by learning in non-social contexts (Shafto, Goodman, & Frank, 2012).

Kushnir, Wellman, and Gelman (2008) show that even children as young as three are sensitive to the knowledgeability of others, and treat intentional actions by knowledgeable agents as more informative about causal structure than actions by ignorant agents. However, children consider both personal and situational constraints on knowledge when evaluating the informativeness of causal interventions. In their study, three- and four-year-olds saw a novel toy that activated in the presence of certain objects. Two actors, one knowledgeable about the toy and one ignorant, each tried to activate the toy with an object. Preschoolers' causal inferences favoured the knowledgeable agent's intervention object only when the agent was allowed to choose it (vs. the child chose it for them), and they could actually apply their knowledge (vs. they where blindfolded when choosing an intervention object). In general, children who demonstrate a stronger understanding of others' mental states are more successful selective causal learners (Legare, Sobel, & Callanan, 2017; D. M. Sobel & Kushnir, 2013).

Finally, if ToM underpins learning about causality, do ToM deficits restrict it? Individuals with a relative inability to process information concerning the mental states of others (e.g. Autism patients) have been shown to make significantly different causal attributions than those individuals skilled at ToM (Blackshaw, Kinderman, Hare, & Hatton, 2001; Kinderman et al., 1998; Randall, Corcoran, Day, & Bentall, 2003) Interestingly, ToM deficits tend to prevent individuals from making situational attributions and increase the probability of agent-oriented causal attributions. Those individuals making relatively many errors on questions measuring ToM ability attribute a significantly greater proportion of causality to causal agents than to causal situational factors (Kinderman, 2001; Kinderman et al., 1998). Channon, Lagnado, Fitzpatrick, Drury, and Taylor (2011) find impaired sensitivity between intentional versus unintentional actions in participants with frontal lesions, with lower causal ratings of intentional acts and higher ratings of unintentional acts relative to control groups. We think these findings correspond to the argument we made in this thesis. People judge causal strength in close alignment with the — inferred or known — mental states of the causal agents. If agents lack significant mental states like intentions or knowledge, they are not preferred as causes over environmental causal factors (Hilton et al., 2016; McClure et al., 2007). If however, the causal observer's or judge's sensitivity towards causal agents' mental states is reduced, it is plausible that they might treat those causal agents as generic causes. Children with deficient ToM abilities do not have impaired perception of causality per se, but struggle to recognise the animacy of artificial animate motion (Congiu, Schlottmann, & Ray, 2010), and in consequence, have difficulties to draw different causal inferences that arise from agent vs. object distinctions

While the studies in this thesis have not been conducted in the area of developmental psychology, we take the large body of research demonstrating the interdependency of ToM and causal learning or inference in development as evidence for the argument we have been aiming to make throughout this thesis: the close connection between ToM and causal reasoning. The fact that both cognitive domains are interrelated in development and learning might lend evidence to their connectedness in higher-level cognition: the fact that we can make sophisticated inferences from causal explanations by ToM, and are sensitive to mental states when judging degrees of causality in agent causation. We began the introduction of this thesis with a summary of the abilities for which causal cognition lays the basis: to explain (Lombrozo, 2010), to make predictions (Einhorn & Hogarth, 1985), to intervene and control (Woodward, 2007), and to make moral judgments (Malle & Nelson, 2003). Because our sense of causality is deeply entrenched in theory of mind, we succeed with these causality-based skills in the social realm, too.

5.2.2 Implications: Artificial Minds

Contemporary formal frameworks of causality have turned a blind eye on the fundamental role of mental states in agent causal reasoning. At the same time, these causal frameworks are used to inform the ever growing use and development of algorithms to reveal causal patterns in large data sets (Pearl, 2019). In order to identify precise causal factors that can provide the basis for intervention and lead to particular behaviours or outcomes, our work suggests that AI should aim to code for 'mental' causes as well. This extension might enable researchers and practitioners to focus on the best mix of interventions for addressing some of today's most critical issues, from climate change to health care.

At the same time, artificial agents present an interesting hybrid case between epistemic and inanimate agents: they are agentive and autonomous, yet often still fall short of human intelligence (Jordan, 2019). With the increasing involvement of algorithmic decision-making in our daily lives, there is already an emerging field of research on normative and psychological theories of responsible machine behaviour (Rahwan et al., 2019). In order to align these theories with our intuitive theories of causality and responsibility, we argue that references to knowledge states of AI systems may be indispensable. However, we currently rest on a very limited conception of what it means for a machine to know a proposition (Bench-Capon, 2014; Rosenschein, 1985). Providing a formalisation of the functional roles of some of the core mental states might resolve this problem (Ashton, 2021; Quillien & German, 2021), but also shed further theoretical insight into our mental taxonomy.

Finally, the rise of AI has not only led to the need of causality and causal inference, but also the need for explanation. Explanations that give insight into algorithmic decision-making – "Explainable AI" – are notoriously difficult to develop (Holzinger, 2018). The integration of Theory of Mind into explanations holds the potential of producing high-quality explanations that are tailored to the beliefs of the listener, in the context of the beliefs of the speaker (Shvo et al., 2020). Endorsing epistemic-based accounts of explanations might help AI systems to produce personalised explanations that take into account users world knowledge and belief revision.

As the work in this thesis suggests, incorporating epistemic states is critical for building better AI systems that are capable of making better causal inferences, taking responsible decisions, justify their actions as well as providing good explanations.

Where do we go from here? We aim to close this thesis by sketching possible fruitful continuations and further developments of the research programme in this thesis.

5.3 Outlook and future research: Theory of Mind in Modal Cognition

In this thesis, we have argued that theory of mind intersects with a seemingly unrelated cognitive domain — the ability for causal judgment. This poses the question whether there are other cognitive abilities that are at first glance unrelated to social cognition, but actually are intimately linked with and draw on ToM in their execution. One such candidate cognitive domain might be the area of modal reasoning, the general ability to reason about alternative possibilities. The impact of normality on causal judgments that has formed a significant part of the research project in this thesis has been demonstrated to impact not only judgments about causation, but a wide-range of non-normative judgments. A large and growing body of research has documented that norm violations influence a variety of intuitive judgments other than causation – such as judgments of *intentional action* (Knobe, 2003), (Kominsky & Phillips, 2019), *freedom* (Young & Phillips, 2011), *happiness* (Phillips, De Freitas, Mott, Gruber, & Knobe, 2017), *doing vs. allowing* (Cushman, Knobe, & Sinnott-Armstrong, 2008), *pro-/con-attitude* ascriptions (Pettit & Knobe, 2009), and *modal* judgments (Knobe & Szabó, 2013).

In the case of causal judgments, abnormal behaviour increases the agent's perceived causal attribution. This impact of the normative status of an agent's action has a variety of different but systematic effects on the judgment types mentioned before. A norm-violating agent is judged as acting more intentionally (Knobe, 2003), having more pro-attitudes towards an outcome (Pettit & Knobe, 2009), being more causal (Icard et al., 2017), being *less* happy (Phillips et al., 2017) and as making (vs. allowing) an outcome to occur, compared to an agent performing the same action but abiding to the norm (Pettit & Knobe, 2009; Phillips et al., 2015).

Theories aiming to explain the influence of norm deviance have all been put forward with reference to causal selection, causal judgments or causal explanations. If the influence of norms however extends to other domains of cognition as well, might these theories also account for the impact of norms on non-causal judgments? In fact, a unified explanation of normality's impact across diverse judgments has been offered (Knobe, 2010; Phillips & Knobe, 2018; Phillips et al., 2015). Similar to the discussed counterfactual theories of causation, this account suggests that the influence of normality on all these judgments —from intentionality to happiness—is driven by people's reasoning about alternatives, or 'possibilities', i.e. modal cognition (Knobe, in press; Phillips & Knobe, 2018).

The work in this thesis as well as in other studies (Kominsky & Phillips, 2019; Samland & Waldmann, 2016) has shown that the effect of norms on judgments of causation are broadly sensitive to agents' epistemic states. Given that causal judgments are just one type of judgments that are sensitive to normality, the question arises whether agents' epistemic states similarly affect the impact of norms across the wide range of different judgments. In a large-scale experimental meta-analysis, Kirfel and Phillips (2021) find evidence for a pervasive impact of ignorance: the impact of norm violations on non-normative judgments depends largely on the agent *knowing* that they were violating a norm when acting.

Current counterfactual theories (Phillips & Knobe, 2018; Phillips et al., 2015) have been aiming to give a unified account explaining the pervasive impact of norms on reasoning about alternatives. However, these frameworks do not yet account for the fact that the impact of normality on people's reasoning about alternatives is diminished when the agent in question is ignorant about their abnormal status. At the same time, the work in this thesis has demonstrated that agent epistemic states do not need to be restricted to norms in order to impact judgements about causation and counterfactual reasoning. What role, then, do epistemic states, broadly construed, play in shaping our reasoning about possibilities?

When humans engage in tasks that require modal cognition (e.g., causal reasoning, moral judgment, judgment about freedom, etc.), it has been suggested that they sample only a small number of specific possible actions (Icard, 2016; Morris et al., 2018; Phillips & Knobe, 2018; Phillips et al., 2019). Recent theoretical advances in modal cognition have offered precise frameworks for how this 'sampling' of alternatives might work. Drawing on the influence of normality (Bear & Knobe, 2017; Phillips et al., 2015), Phillips et al. (2019) for example suggest that our minds adaptively sample a small set of possible actions that are relevant for a specific task and, from within the relevant space, consider only those actions that are statistically likely and and morally valuable.

A promising continuation of this line of research in this thesis is hence to investigate which role theory of mind occupies in judgments that rely on modal reasoning about third-party agents' actions. How does what an agent believes and knows affect the alternatives we imagine for their action? One obvious hypothesis is that our theory of the agent's mind restricts the set of possible actions we can sample from.

Put simply, only an action that the agent in focus considers possible is an action that *we* as reasoners will consider possible when reasoning about what the agent could have done differently. In that sense, the set of possible actions that we as reasoners consider are restricted by the causal agent's model of the world — what the agent knows about rules, norms, but also their causal knowledge or knowledge about facts. Within the agent's view about what's possible, we might then be biased towards sampling normal and probable actions (Phillips et al., 2019). Integrating theory of mind into accounts of adaptive sampling of possible actions raises a variety of interesting theoretical and empirical questions. Is sampling of normal actions dependent on what we as observer consider as normal, or is this again affected by ToM, the agent's perspective of what's normal? Finally, the work in this thesis has mainly focused on epistemic, i.e. knowledge-related states. It still remains to be investigated if other kind of mental states, for example intentions, beliefs or desires, also influence how we think about alternatives.

While first agent-specific accounts of modality have been developed in semantics (Maier, 2015; Mandelkern, Schultheis, & Boylan, 2017), a unified account for modal reasoning integrating agent mental states still needs to be worked out. We envisage the study of theory of mind in modal cognition as a fruitful continuation of the research programme sketched in this thesis. Counterfactual theories drawing on the tradition of Lewis (1986)'s possible worlds theory often assume that people consider the "closest" possible world when thinking about alternatives (De Brigard, Henne, & Stanley, 2021). The work in this research suggests that we might need to reconsider Lewis' metrics of "close neighbours". What the agent could have done differently can only be evaluated in the closest possible world *the agent knows about*.

Ultimately, judgments of causation might not be the only cognitive domain that draws on counterfactual reasoning. The work in this thesis suggests that theory of mind intersects with how we think about alternatives, and the inferences we draw from them. Future research will show how such an integration of theory of mind on the one hand and reasoning about 'possible worlds' in broader human cognition on the other hand might look like.

Appendix A

Appendix Chapter 2

A.1 Comprehension Check Questions Experiment 1

Participants answered ten comprehension check questions in total, five for each *'normality'* condition.

[Part 1]Tom and Ben ...

1. - work in the same office.
   - work in separate offices.

2. Because of their office situation, Tom and Ben ...

   - know each other well.
   - do not know each other, and have never met or seen each other.

3. Because of their office situation, Tom and Ben ...

   - know what the other person is doing during the day.
   - do not know what the other person is doing during the day.

4. The use of how many microwaves does it take to produce a power failure on Friday?

   - one microwave.
   - two microwaves.

[Part 2]

5. Who used a microwave frequently (rather than infrequently) this week? (Multiple answers possible)

    - Tom.
    - Ben.

A.2 Comprehension Check Questions Experiment 2

[Part 1]Tom and Ben ...

1. - work in the same office.
    - work in separate offices.

2. Because of their office situation, Tom and Ben ...

    - know each other well.
    - do not know each other, and have never met or seen each other.

3. Because of their office situation, Tom and Ben ...

    - know what the other person is doing during the day.
    - do not know what the other person is doing during the day.

4. The use of how many microwaves does it take to produce a power failure on Friday?

    - one microwave.
    - two microwaves.

[Part 2]

5. Who used a microwave frequently (rather than infrequently) this week? (Multiple answers possible)

    - Tom.
    - Ben.

A.3 Comprehension Check Questions Experiment 3

[Part 1]From Monday to Thursday, Tom and Ben ...

1. - work in the same office.
   - work in separate offices.

2. Because of their office situation, Tom and Ben ...

   - know each other well.
   - do not know each other, and have never met or seen each other.

3. Because of their office situation, Tom and Ben ...

   - know what the other person is doing during the day.
   - do not know what the other person is doing during the day.

4. The use of how many microwaves does it take to produce a power failure on Friday?

   - one microwave.
   - two microwaves.

   [Part 2]

5. Who used a microwave frequently (rather than infrequently) this week? (Multiple answers possible)

   - Tom.
   - Ben.

A.4 Comprehension Check Questions Experiment 4

[Part 1]Tom and Ben ...

1. - work in the same office.
   - work in separate offices.

2. Because of their office situation, Tom and Ben ...

    - know each other well.
    - do not know each other, and have never met or seen each other.

3. Because of their office situation, Tom and Ben ...

    - know what the other person is doing during the day.
    - do not know what the other person is doing during the day.

4. The use of how many microwaves does it take to produce a power failure on Friday?

    - one microwave.
    - two microwaves.

[Part 2]

5. Who used a microwave frequently (rather than infrequently) this week? (Multiple answers possible)

    - Tom.
    - Ben.

Appendix B

Appendix Chapter 3

B.1 Scenarios

B.1.1 Scenario “Garden”

Please imagine the following scenario:

(Part 1)

“Bob is a gardener in a local botanical garden and takes care of a very delicate type of rose, the China rose. Bob regularly fertilizes the roses with the fertilizer “Vitax” in order to keep them alive and healthy. The botanical garden supplies all gardeners who work in the botanical garden with gardening tools, chemicals and fertilizer. Normally, fertilizers do not harm delicate roses.

The garden manager has recently started to order an additional fertilizer, “Nutrit”, that is cheaper than “Vitax”. “Nutrit” is as effective as “Vitax”, but also has a negative effect. It harms delicate rose types such as the China rose.

Because the new fertilizer has only recently been ordered, Bob does not know that it harms delicate rose types such as the China rose.”

(Part 2)

“One day, Bob takes care of the China roses in the botanical garden. Bob does not know that the new fertilizer “Nutrit” harms China roses.

Bob fertilizes the China roses with the fertilizer “Nutrit”. As a result of the fertilization, the China roses die.”

B.1.2 Scenario “Bakery”

Please imagine the following scenario:

(Part 1)

“Anne is a baker who works for the local bakery in town. The bakery offers a variety of pastries, including nut allergy friendly cakes, muffins and cookies. Anne uses flour of the brand “Green Farms” when baking. The bakery provides all necessary baking products including flour. Normally, flour does not contain traces of nuts.

The bakery manager has recently started to order an additional flour brand, “Homestead”, that is cheaper than “Green Farms”. “Homestead” is of the same quality as “Green Farms”, but differs in one aspect. It contains traces of hazelnuts.

Because the new flour has only recently been ordered, Anne does not know that it contains traces of hazelnuts.”

(Part 2)

“One day, Anne is baking an allergy friendly cake for a customer with nut allergy. Anne does not know that the “Homestead” flour includes traces of hazelnuts.

Anne uses the “Homestead” flour for the cake. As a result of using this flour, the customer suffers from an allergic reaction.”

For the scenarios of Experiment 2-4, please visit `https://github.com/LaraKirfel/EpistemicInterventions`.

B.2 Between Subject Results

B.2.1 Experiment 1, Between Subject Factor Analysis

B.2.1.1 Causal Ratings

A Knowledge × Scenario ANOVA indicated a sign. effect for the factor Knowledge, $F(1) = 92.5$; $p < .001$. Ratings in the *knowledge* condition (*M* = 6.19, *SD* = 1.25, 95% CI [5.87, 6.52]) were higher than in the *ignorance* condition (*M* = 4.31, *SD* = 2.13, 95% CI [3.80, 4.83].). The ANOVA also indicated a sign. effect for the interaction between ‘Knowledge’ × ‘Scenario’, $F(1) = 4.57$; $p = .01$. There was no sign. difference between the knowing agent (*M* = 6.05, *SE* = .40, 95% CI [5.28, 6.83]) and the ignorant agent in the bakery scenario,(*M* = 5.37, *SE* = .40, 95% CI [4.59, 6.14]), $t(115) = 1.24$; $p = .22$.

B.2.1.2 Counterfactual Responses

. Addition of the knowledge factor significantly improves the fit for predicting people's counterfactual responses, χ^2(-3) = 52.48; *p* ¡ .001, R^2= .17]. When the agent's epistemic state changes from knowledge to ignorance, people are less likely to imagine a counterfactual change that concerns the agent's action (75% vs. 30%) (*b* = -1.00, *OR* = -.37, *SE* = .44, *z* = -2.26, $p < .001$). In the "knowledge" condition, no responses concerning epistemic states were given.

B.2.2 Experiment 2, Between Subject Factor Analysis

B.2.2.1 Causal Ratings

An ANOVA indicated that ignorance was a significant factor for causal ratings, $F(1)$ = 23.85; $p < .001$. People's causal ratings were lower when the agent's ignorance was caused externally (*M* = 3.68, *SD* = 2.25, 95% CI [3.17, 4.20]) rather than by choice (*M* =5.27, *SD* = 1.70 95%, CI [4.89, 5.65]). There was no significant effect of scenario $F(1) = 0.32$; $p = .73$, nor an interaction between scenario and ignorance, $F(1) = 0.81$; $p = .44$.

B.2.2.2 Counterfactual Responses

Addition of the knowledge factor significantly improves the fit for predicting people's counterfactual responses, χ^2(-4) = 53.87; $p < .001$, R^2= .17]. The change from self-caused to externally caused ignorance people increases the relative log odds to indicate an externally caused epistemic change over an action change (*b* = 2.14, *OR* = 8.48, *SE* = .67, *z* = 3.20, *p* = .001), people are more likely to imagine an epistemic state change that is caused by external or other factors in the self (4% vs. 52%).

B.2.2.3 Counterfactual Responses: Subgroup Analysis

A subgroup analysis showed that type of counterfactual response chosen predicted people's causal ratings in addition to the ignorance condition, $F(1) = 6.22$; $p = .01$. Those people in the "externally caused ignorance" condition who imagined a self-caused epistemic change gave a higher causal rating (*M* =4.63, *SD* = 2.15, 95% CI [3.73, 5.53]) than those who imagined an externally caused epistemic change (*M* =3.32, *SD* = 2.19, 95% CI [2.62, 4.03]), *t*(121) = 2.63, *p* = .01. In the "self-caused ignorance" condition, there is no difference in ratings between participants who stated a external (*M* =5.67, *SD* = 1.15, 95% CI [4.36, 6.97]) vs self-caused epistemic change (*M* =5.47, *SD* = 1.51, 95% CI [5.10, 5.85]), *t*(121) = -0.17, *p* = .86].

B.2.2.4 Knowledge Ratings

An ANOVA indicated that ignorance was a significant factor for knowledge ratings, $F(1)$ = 56.18; $p < .001$. People's knowledge ratings were lower when the agent's ignorance was caused externally (*M* = 3.65, *SD* = 2.20, 95% CI [3.14, 4.16]) rather than by choice (*M* = 6.04, *SD* = 1.56, 95%, CI [5.66, 6.42]). There was a significant effect of scenario $F(2)$ = 3.78; *p* = .02, but no interaction effect, $F(2)$ = 0.38; *p* = .68. Ratings in the "bakery" scenario were lower (*M* = 4.36, *SD* = 2.31, 95% CI [3.82, 4.89]) than in the "hospital" scenario (*M* = 5.60, *SD* = 2.09, 95% CI [4.97, 6.24]), *t*(147) = -2.86, *p* = .01.

B.2.2.5 Blame Ratings

An ANOVA indicated that ignorance was a significant factor for blame ratings, $F(1)$ = 98.76; *p* ¡ .001. People's blame ratings were lower when the agent's ignorance was caused externally (*M* = 2.93, *SD* = 1.99, 95% CI [2.42, 3.44]) rather than by choice (*M* = 5.76, *SD* = 1.41, 95%, CI [5.38, 6.14]). There was a significant effect of scenario $F(2)$ = 4.90; *p* = .01, but no interaction effect, $F(2)$ = 0.19; *p* = .82. Ratings in the "hospital" scenario were higher (*M* = 5.17, *SD* = 2.14, 95% CI [4.54, 5.81]) than in the garden scenario (*M* = 3.93, *SD* = 2.17, 95% CI [3.30, 4.57]), *t*(144) = -2.83, *p* = .01, and the bakery scenario, (*M* = 4.01, *SD* = 2.19, 95% CI [3.49, 4.55]), *t*(144) = -2.61, *p* = .03.

B.2.3 Experiment 3, Between Subject Factor Analysis

B.2.3.1 Causal Rating

The ignorance factor was not a significant predictor for participants' causal responses, *between-subject contrast*: $F(1) = 0.03$; $p = .87$. There was no significant effect of scenario, $F(2) = 1.24$; $p = .29$ and no interaction between ignorance and scenario ($p = .87$).

B.2.3.2 Counterfactual Responses

A model with the ignorance condition as predictor provided a significant fit for people's counterfactual responses, χ^2(-5) = 29.17; $p < .001$ R^2= .11. Changing the ignorance condition from "many actions" to "few actions" significantly increases the log odds of a response indicating an epistemic state changed caused by an alternative action of the agent "... by other" ($b = 2.30$, $OR = 10$, $SE = 0.99$, $z = 2.33$, $p =$ .01). (4% vs. 15%).

B.2.3.3 Counterfactual Responses: Subgroup Analysis

Adding a predictor "counterfactual response type" to a model already including ignorance condition did not provide a better fit for people's causal judgments, $F(1)$ = 0.51; $p = .48$.

B.2.3.4 Knowledge & Blame Ratings

Agreement ratings with the statement that the agent could have known about the harmful properties of their action were not influenced by the number of actions necessary for knowledge, $F(1) = 0.52$; $p = .47$]. Blame ratings were also not influenced by the ignorance condition factor $F(1) = 0.55$; $p = .46$]. In addition to the epistemic action condition, knowledge rating was a significant predictor for people's causal judgements, $F(1) = 32.70$; $p < .001$ ($b = 0.36$, $SE = .10$, $t = 3.82$) and blame ratings, $F(1) = 53.37$; $p < .001$ ($b = 0.28$, $SE = .09$, $t = 3.20$).

B.2.4 Experiment 4, Between Subject Factor Analysis

B.2.4.1 Causal Rating

The “information seeking” factor was a significant predictor for participants’ causal responses, $F(1) = 12.00$; $p < .001$. People judged the agent to be less of a cause (b = -.71, SE = .22, t = -3.25) when the agent read the e-mail with the missing information (M = 3.06, SD = 2.20, 95% CI [2.55, 3.56]) than if they did not (M = 4.34, SD = 1.97, 95%, CI [3.85, 4.84]).

B.2.4.2 Counterfactual Responses

The information acquisition condition significantly predicted people’s counterfactual responses, χ^2(-4) = 65.49; $p < .001$, R^2= .20. When the agent did not read the e-mail, people were less likely to indicate a change that consisted in the addition of *just* the missing information in the e-mail (b = -1.80, OR = -.16, SE = .91, z = -1.96, $p < .001$) (6% vs. 60%).

B.2.4.3 Counterfactual Responses: Subgroup Analysis

Adding “counterfactual response type” as a predictor to a model including the “epistemic action” factor significantly improved the fit of the model for causal judgments, $F(1) = 15.34$; $p < .001$]. In the “doesn’t read e-mail” condition, people who imagined the agent to perform an alternative action in order acquire knowledge gave higher causal ratings (M = 5.50, SD = 1.71, 95% CI [4.66, 6.34]) than those who indicated the e-mail could have had the relevant information and the agent could have read it, (M =3.36, SD = 2.03, 95% CI [2.71, 4.55]), $t(87)$ = -3.04, $p < .01$. The difference in causal ratings between “by info” (M = 2.38, SD = 1.75, 95% CI [1.85, 2.91]) and “by other action of agent” (M = 5.25, SD = 1.91, 95% CI [4.17, 6.33]) responders was also significant in the “reads e-mail” condition, $t(87)$ = -4.84, $p <$ .001.

B.2.4.4 Knowledge and Blame Rating

Information-seeking behaviour significantly predicted modal judgments about the agent’s epistemic state, $F(1) = 4.38$; $p = .03$, as well as blameworthiness for ignorance, $F(1) = 18.42$; $p < .001$. The agent who did not read the e-mail containing missing information was judged to could have known about the relevant information to a slightly greater extent (*M* = 3.80, *SD* = 2.17, 95% CI [3.31, 4.30]) and to blame slightly more for their ignorance (*M* = 3.85, *SD* = 2.06, 95% CI [3.36, 4.35]) than the information-seeking agent ("Could have known": *M* = 2.99, *SD* = 2.29, 95% CI [2.48, 3.50]; "Blame": *M* = 2.39, *SD* = 1.96, 95% CI [1.88, 2.90])

Adding knowledge rating as a factor significantly improved a model that already contained the "epistemic action" condition for people's causal ratings, $F(1)$ = 31.27, $p < .001$, (*b* = 0.56, *SE* = .11, *t* = 4.98) as well as blame ratings, $F(1)$ = 72.96, $p < .001$ (*b* = 0.76, *SE* = .09, *t* = 8.14).

References

Ahn, W.-K., & Bailenson, J. (1996). Mechanism-based explanations of causal attribution: An explanation of conjunction and discounting effect. *Cognitive Psychology*, *31*, 82–123.

Ahn, W.-K., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 199–225). Cambridge, MA: Cambridge University Press.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action control* (pp. 11–39). Springer.

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

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*, 179–186.

Alicke, M. D., & Rose, D. (2012a). Culpable control and causal deviance. *Journal of Personality and Social Psychology Compass*, *6*, 723–725.

Alicke, M. D., & Rose, D. (2012b). Culpable control and causal deviance. *Social and Personality Psychology Compass*, *6*(10), 723–735.

Alicke, M. D., Rose, D., & Bloom, D. (2012). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696.

Alwin, D. F. (1973). Making inferences from attitude-behavior correlations. *Sociometry*, 253–278.

Ashton, H. (2021). Definitions of intent suitable for algorithms. *arXiv preprint arXiv:2106.04235*.

Baillargeon, R. (1995). Physical reasoning in infancy. *The cognitive neurosciences*, 181–204.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Baker, C. L. (2012). *Bayesian theory of mind: Modeling human reasoning about beliefs, desires, goals, and social relations* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Bandura, A. (1962). Social learning through imitation.

Barbero, F., Schulz, K., Smets, S., Velázquez-Quesada, F. R., & Xie, K. (2020). Thinking about causation: A causal language with epistemic operators. In *International workshop on dynamic logic* (pp. 17–32).

Baron-Cohen, S. (1996). Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition*, *3*(1), 39–60.

Barton, K., & Barton, M. K. (2015). Package 'mumin'. *Version*, *1*, 18.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* Oxford university press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version*, *2*(1), 74.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, *194*, 104057.

Bear, A., & Knobe, J. (2017, October). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37. doi: 10.1016/j.cognition.2016.10.024

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(1), 153–164.

Bench-Capon, T. J. (2014). *Knowledge representation: An approach to artificial intelligence* (Vol. 32). Elsevier.

Blackshaw, A. J., Kinderman, P., Hare, D. J., & Hatton, C. (2001). Theory of mind, causal attribution and paranoia in asperger syndrome. *Autism*, *5*(2), 147–163.

Bonnefon, J.-F. (2007). Reasons to act and the mental representation of consequentialist aberrations. *Behavioral and Brain Sciences*, *30*(5-6), 453.

Bonnefon, J.-F., Zhang, J., & Deng, C. (2007). L'effet des justifications sur le regret est-il direct ou indirect? *Revue internationale de psychologie sociale*, *20*(2), 131–145.

Bosco, F. M., Tirassa, M., & Gabbatore, I. (2018). Why pragmatics and theory of mind do not (completely) overlap. *Frontiers in Psychology*, *9*, 1453.

Bramley, N., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In *Cogsci.*

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive psychology*, *105*, 9–38.

Brandstätter, V., Lengfelder, A., & Gollwitzer, P. M. (2001). Implementation intentions and efficient action initiation. *Journal of personality and social psychology*, *81*(5), 946.

Branscombe, N. R., Wohl, M. J., Owen, S., Allison, J. A., & N'gbala, A. (2003). Counterfactual thinking, blame assignment, and well-being in rape victims. *Basic and Applied Social Psychology*, *25*(4), 265–273.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.

Brickman, P., Ryan, K., & Wortman, C. B. (1975). Causal chains: Attribution of responsibility as a function of immediate and prior causes. *Journal of Personality and Social Psychology*, *32*(6), 1060–1067.

Bruckmüller, S., Hegarty, P., Teigen, K. H., Böhm, G., & Luminet, O. (2017, Jun). When do past events require explanation? insights from social psychology. *Memory Studies*, *10*(3), 261–273.

Buss, A. R. (1978). Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology*, *36*(11), 1311–1321.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality.

Psychological Review, *90*(2), 105.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135–157.

Chang, L. J., & Sanfey, A. G. (2009). Unforgettable ultimatums? expectation violations promote enhanced social memory following economic bargaining. *Frontiers in Behvarioal Neuroscience*, *3*.

Channon, S., Lagnado, D., Fitzpatrick, S., Drury, H., & Taylor, I. (2011). Judgments of cause and blame: Sensitivity to intentionality in asperger's syndrome. *Journal of Autism and Developmental Disorders*, *41*(11), 1534–1542.

Chee, C. S., & Murachver, T. (2012). Intention attribution in theory of mind and moral judgment. *Psychological Studies*, *57*(1), 40–45.

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

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*(3), 439–477.

Chisholm, R. M. (1976). The agent as cause. In *Action theory* (pp. 199–211). Springer.

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

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*(5), 461–480.

Clark, C. J., Baumeister, R. F., & Ditto, P. H. (2017). Making punishment palatable: Belief in free will alleviates punitive distress. *Consciousness and Cognition*, *51*, 193–211.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F.

(2014). Free to punish: A motivated account of free will belief. *Journal of personality and social psychology*, *106*(4), 501.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clarke, R., Shepherd, J., Stigall, J., Waller, R. R., & Zarpentine, C. (2015). Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology*, *28*(2), 279–293.

"Climate crisis 'unequivocally' caused by human activities, says IPCC report". (2021, August 9).

The Guardian. Retrieved from `https://www.theguardian.com/environment/2021/aug/09/climate-crisis-unequivocally-caused-by-human-activities-says-ipcc-report`

Coffa, J. A. (1974). Hempel's ambiguity. *Synthese*, *28*, 141-163.

Collins, J. (2000, Apr). Preemptive prevention. *The Journal of Philosophy*, *97*(4), 223. Retrieved from `http://dx.doi.org/10.2307/2678391` doi: 10.2307/2678391

Congiu, S., Schlottmann, A., & Ray, E. (2010). Unimpaired perception of social and physical causality, but impaired perception of animacy in high functioning children with autism. *Journal of autism and developmental disorders*, *40*(1), 39–53.

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, Mar). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. Retrieved from `http://dx.doi.org/10.1371/journal.pone.0057410` doi: 10.1371/journal.pone.0057410

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 447–458.

Cummings, L. (2015a). Pragmatic disorders and theory of mind. In *The cambridge* *handbook of communication disorders* (pp. 559–577). Cambridge University Press.

Cummings, L. (2015b). Theory of mind in utterance interpretation: the case from clinical pragmatics. *Frontiers in Psychology*, *6*, 1286.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97–103.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, *108*(1), 281–289.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.

Danks, D. (2013). Functions and cognitive bases for the concept of actual causation. *Erkenntnis*, *78*(S1), 111–128. Retrieved from `https://doi.org/10.1007%2Fs10670-013-9439-2` doi: 10.1007/s10670-013-9439-2

Danks, D. (2017). Singular causation. In M. Waldmannn (Ed.), *The oxford handbook of causal reasoning* (pp. 201–215). Oxford University Press.

Danks, D., Rose, D., & Machery, E. (2014). Demoralizing causation. *Philosophical Studies*, *171*(2), 251–277. Retrieved from `https://doi.org/10.1007%2Fs11098-013-0266-8` doi: 10.1007/s11098-013-0266-8

Danks, D., Rose, D., & Machery, E. (in press). Demoralizing causation. *Philosophy and Phenomenological Research*.

Danner, U. N., Aarts, H., & de Vries, N. K. (2008). Habit vs. intention in the prediction of future behaviour: The role of frequency, context stability and mental accessibility of past behaviour. *British Journal of Social Psychology*, *47*(2), 245–265.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*(4), 324–336.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700.

Davis, Z. J., Allen, K. R., & Gerstenberg, T. (2021). Who went fishing? inferences from social evaluations.

De Brigard, F., Henne, P., & Stanley, M. L. (2021). Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility. *Cognition*, *209*, 104574.

Dehghani, M., Iliev, R., & Kaufmann, S. (2007). Effects of fact mutability in the interpretation of counterfactuals. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55–85.

Dickson, V., & Theobald, J. (2011). Criminal law. In *Graduate diploma in law.* Guildford: The College of Law.

Dolan, P., Elliott, A., Metcalfe, R., & Vlaev, I. (2012). Influencing financial behavior: From changing minds to changing contexts. *Journal of Behavioral Finance*, *13*(2), 126–142.

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, *33*(1), 264–277.

Dowe, P. (2001, jun). A counterfactual theory of prevention and "causation" by omission. *Australasian Journal of Philosophy*, *79*(2), 216–226. Retrieved from `http://dx.doi.org/10.1080/713659223` doi: 10.1080/713659223

Driver, J. (2008). Attributions of causation and moral responsibility.

Einhorn, H. J., & Hogarth, R. M. (1985). Prediction, diagnosis, and causal thinking in forecasting. In *Behavioral decision making* (pp. 311–328). Springer.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3.

Engelmann, N., & Waldmann, M. R. (2021). A causal proximity effect in moral judgment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Epstein, S. (1979). The stability of behavior: I. on predicting most of the people much of the time. *Journal of personality and social psychology*, *37*(7), 1097.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and social psychology review*, *12*(2), 168–192.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, *14*(3), 219–250.

Fausey, C. M., & Boroditsky, L. (2007). Language changes causal attributions about agents and objects. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Fazelpour, S. (2020, Jun). Norms in counterfactual selection. *Philosophy and Phenomenological Research*. Retrieved from `http://dx.doi.org/10.1111/phpr.12691` doi: 10.1111/phpr.12691

Fenton-Glynn, L. (2017). A proposed probabilistic extension of the halpern and pearl definition of 'actual cause'. *The British Journal for the Philosophy of Science*, *68*(4), 1061–1124.

Fernández, C. (2013). Mindful storytellers: Emerging pragmatics and theory of mind development. *First Language*, *33*(1), 20–46.

Ferrante, D., Girotto, V., Stragà, M., & Walsh, C. (2013). Improving the past and the future: A temporal asymmetry in hypothetical thinking. *Journal of Experimental Psychology: General*, *142*(1), 23.

Fillon, A., Kutscher, L., & Feldman, G. (2020). Impact of past behaviour normality: meta-analysis of exceptionality effect. *Cognition and Emotion*, 1–21.

Fillon, A., Lantian, A., Feldman, G., & N'gbala, A. (2019). Exceptionality effect in agency: Exceptional choices attributed higher free will than routine.

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, *39*(2), 75–85.

FitzPatrick, W. J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics*, *118*(4), 589–613.

FitzPatrick, W. J. (2017). Unwitting wrongdoing, reasonable expectations, and blameworthiness. *Responsibility: The epistemic condition*, 29–46.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of*

Philosophy, *71*(1), 5-19.

Friedman, W. J. (2001). The development of an intuitive understanding of entropy. *Child Development*, *72*(2), 460–473.

Frith, C., & Frith, U. (2005). Theory of mind. *Current biology*, *15*(17), R644–R645.

Gavanski, I., & Wells, G. L. (1989). Counterfactual processing of normal and exceptional events. *Journal of Experimental Social Psychology*, *25*(4), 314–325.

Gergely, G., Bekkering, H., & Király, I. (2002). Developmental psychology: Rational imitation in preverbal infants. *Nature*, *415*(6873), 755.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gerstenberg, T., Goodman, N. D., Lagnado, D., & Tenenbaum, J. (2020, Mar). *A counterfactual simulation model of causal judgments for physical events.* PsyArXiv. doi: 10.31234/osf.io/7zj94

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015a). How, whether, why: Causal judgments as counterfactual contrasts. In *Cogsci.*

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015b). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 788–793). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599.

Gerstenberg, T., & Icard, T. F. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General.*

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. (2014). Attributing responsibility: Actual and counterfactual worlds. In J. Knobe, T. Lombrozo, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 1, pp. 91–130). Oxford University Press.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from https://doi.org/10.1177%2F0956797617713053 doi: 10.1177/0956797617713053

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018a). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018b, August). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122-141. doi: 10.1016/j.cognition.2018.03.019

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 409–414). Austin, TX: Cognitive Science Society.

Gibbons, J. (2001). Knowledge in action. *Philosophy and Phenomenological Research*, *62*(3), 579–600.

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

Glautier, S. (2008). Recency and primacy in causal judgments: Effects of probe question and context switch on latent inhibition and extinction. *Memory & cognition*, *36*(6), 1087–1093.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., … Zhang, J. (2010). Actual causation: a stone soup essay. *Synthese*, *175*(2), 169–192.

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. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2759–2764).

Goodman, N. D., & Frank, M. C. (2016, nov). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. Retrieved from `https://doi.org/10.1016%2Fj.tics.2016.08.005` doi: 10.1016/j.tics.2016.08.005

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*. Retrieved from `http://www.mit.edu/~ast/papers/implicature-topics2013.pdf`

Goodwin, G. P. (2014). How complete is the path model of blame? *Psychological Inquiry*, *25*(2), 215–221.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Green, P., & MacLeod, C. J. (2016). Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts.* New York: Wiley.

Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020a). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, 1069. Retrieved from `https://www.frontiersin.org/article/10.3389/fpsyg.2020.01069` doi: 10.3389/fpsyg.2020.01069

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020b). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, 1069.

Guglielmo, S., & Malle, B. F. (2017). Information-acquisition processes in moral judgments of blame. *Personality and Social Psychology Bulletin*, *43*(7), 957–971.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one*, *14*(3), e0213544.

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: causal bayes nets as rational models of everyday causal reasoning. *Synthese*, *189*(1), 17–28.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, 86–100.

Hall, N., Paul, L. A., et al. (2003). *Causation and pre-emption.* Philosophy of science today. New York: Oxford University Press.

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. (2016). *Actual causality*. MIT Press.

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. (2015). Graded causation and defaults. *British*

Journal for the Philosophy of Science, *66*, 413–457.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hanna, E., & Meltzoff, A. N. (1993). Peer imitation by toddlers in laboratory, home, and day-care contexts: Implications for social learning and memory. *Developmental psychology*, *29*(4), 701.

Happé, F., & Loth, E. (2002). 'theory of mind'and tracking speakers' intentions. *Mind & Language*, *17*(1-2), 24–36.

Harinen, T. (2017, jul). Normal causes for normal effects: Reinvigorating the correspondence hypothesis about judgments of actual causation. *Erkenntnis*. Retrieved from `https://doi.org/10.1007%2Fs10670-017-9876-4` doi: 10.1007/s10670-017-9876-4

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

Hawthorne, J., & Stanley, J. (2008). Knowledge and action. *The Journal of Philosophy*, *105*(10), 571–590.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, *51*(6), 358–374.

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

Heider, F. (1983). *The psychology of interpersonal relations*. Psychology Press.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Hemmatian, B., & Sloman, S. A. (2018). Community appeal: Explanation without information. *Journal of Experimental Psychology: General*, *147*(11), 1677.

Henne, P., Bello, P., Khemlani, S., & De Brigard, F. (2019). Norms and the meaning of omissive enabling conditions.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2020). Counterfactual thinking and recency effects in causal judgment.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

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

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2019, Dec). *Norms affect prospective causal judgments.* OSF Preprints. Retrieved from `osf.io/2nwb4` doi: 10.31219/osf.io/2nwb4

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science*, *45*(1), e12931.

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

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Brighton, UK: Harvester Press.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81.

Hilton, D. J. (1996, Nov). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273–308.

Hilton, D. J., & Jaspars, J. M. (1987). The explanation of occurrences and non-occurrences: A test of the inductive logic model of causal attribution. *British Journal of Social Psychology*, *26*(3), 189–201.

Hilton, D. J., McClure, J., & Moir, B. (2016). Acting knowingly: effects of the agent's awareness of an opportunity on causal attributions. *Thinking & Reasoning*, *22*(4), 461–494.

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. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

Hitchcock, C. (2017). Actual causation: What's the use? In *Making a difference: Essays on the philosophy of causation.* Oxford University Press.

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

The holy bible. (n.d.). New International Version®, NIV® Copyright© 1973, 1978, 1984, 2011 by Biblica.

Holzinger, A. (2018). From machine learning to explainable ai. In *2018 world symposium on digital intelligence for systems and machines (disa)* (pp. 55–66).

Hume, D. (1748/1975). *An enquiry concerning human understanding*. Oxford University Press.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

Icard, T., & Knobe, J. (2016). Causality, normality, and sampling propensity. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 800–805). Austin, TX: Cognitive Science Society.

Icard, T., Kominsky, J., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

"In 2005, Helios flight 522 crashed into a Greek hillside. Was it because one man forgot to flip a switch?". (2020, September 19). *The Guardian*. Retrieved from `https://www.theguardian.com/world/2020/sep/19/in-2005-helios-flight-522-crashed-into-a-greek-hillside-was-it-because-one-man-forgot-to-flip-a-switch`

"invertedparadoxxx". (2019). Romeo and juliet would still be alive if he had checked her pulse [online forum post]. *Reddit Forum.* Retrieved from `https://www.reddit.com/r/Jokes/comments/bnhxyo/romeo_and_juliet_would_still_be_alive_if_he_had`

Jackson, F. (1977, may). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, *55*(1), 3–21. Retrieved from `http://dx.doi.org/10.1080/00048407712341001` doi: 10.1080/00048407712341001

Jackson, F. (1996). Mental causation. *Mind*, *105*(419), 377–413.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 785. Retrieved from `https://doi.org/10.1016%2Fj.tics.2016.08.007` doi: 10.1016/j.tics.2016.08.007

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive psychology*, *77*, 42–76.

Johnson-Laird, P., & Khemlani, S. (2017). Mental models and causation. *Oxford handbook of causal reasoning*, 1–42.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in cognitive sciences*, *19*(4), 201–214.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. In *Advances in experimental social psychology* (Vol. 2, pp. 219–266). Elsevier.

Jones, E. E., Davis, K. E., & Gergen, K. J. (1961). Role playing variations and their informational value for person perception. *The Journal of Abnormal and Social Psychology*, *63*(2), 302.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of experimental social psychology*, *3*(1), 1–24.

Jordan, M. I. (2019). Artificial intelligence—the revolution hasn't happened yet.

Harvard Data Science Review, *1*(1).

Juhos, C., Quelhas, A. C., & Byrne, R. M. (2015). Reasoning about intentions: Counterexamples to reasons for actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 55.

Kahneman, D. (2014). Varieties of counterfactual thinking. In *What might have been* (pp. 387–408). Psychology Press.

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

Kamide, Y. (2012, July). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71. doi: 10.1016/j.cognition.2012.03.001

Kasimatis, M., & Wells, G. L. (1995). Individual differences in counterfactual thinking. *What might have been: The social psychology of counterfactual thinking*, 81–101.

Keil, F. (2006). Explanation and understanding. *Annual review of psychology*, *57*, 227.

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences*, *3*(12), 461–468.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation.*

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

Kerwer, M., Rosman, T., Wedderhoff, O., & Chasiotis, A. (2021). Disentangling the process of epistemic change: The role of epistemic volition. *British Journal of Educational Psychology*, *91*(1), 1–26.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & cognition*, *46*(8), 1344–1359.

Khemlani, S. S., Byrne, R. M., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive science*, *42*(6), 1887–1924.

Kim, J. (1974). Causes and counterfactuals. *The Journal of Philosophy*, *70*(17), 570–572.

Kim, J. (2007). Causation and mental causation. *Contemporary debates in philosophy of mind*, 227–242.

Kinderman, P. (2001). *Changing causal attributions.* American Psychological Association.

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204.

Kirfel, L., Icard, T., & Gerstenberg, T. (2020, May). *Inference from explanation.* PsyArXiv. Retrieved from psyarxiv.com/x5mqc doi: 10.31234/osf.io/x5mqc

Kirfel, L., & Lagnado, D. (2020, Aug). *Causal judgments about atypical actions are influenced by agents' epistemic states.* PsyArXiv. Retrieved from psyarxiv.com/yvstb doi: 10.31234/osf.io/yvstb

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027721001402 doi: https://doi.org/10.1016/j.cognition.2021.104721

Kirfel, L., & Lagnado, D. A. (2017). "Oops, I did it again." The impact of frequent behaviour on causal judgement. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2420–2425). Austin, TX: Cognitive Science Society.

Kirfel, L., & Lagnado, D. A. (2018). Statistical norm effects in causal cognition. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 615–620). Austin, TX: Cognitive Science Society.

Kirfel, L., & Lagnado, D. A. (2019). I know what you did last summer (and how often). epistemic states and statistical normality in causal judgments. *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*

Kirfel, L., & Phillips, J. (2021). The impact of ignorance beyond causation: An experimental meta-analysis. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, *18*(4), 513–549.

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, *169*, 139–146.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

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), 315–329. doi: 10.1017/S0140525X10000907

Knobe, J. (in press). *Morality and possibility.* The Oxford Handbook of Moral Psychology. Oxford: Oxford University Press.

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., & Shapiro, S. (2021). Proximate cause explained. *The University of Chicago Law Review*, *88*(1), 165–236.

Knobe, J., & Szabó, Z. G. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics*, *6*, 1–1.

Kominsky, J. F., Gerstenberg, T., Pelz, M., Singmann, H., Sheskin, M., & Keil, F. (2019). The trajectory of counterfactual simulation in development. In *Cogsci* (pp. 2044–2050).

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, *43*(11), e12792.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kominsky, J. F., Phillips, J., Knobe, J., Gerstenberg, T., & Lagnado, D. A. (2014). Causal supersession. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 761–766). Austin, TX: Cognitive Science Society.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological science*, *14*(5), 402–408.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental psychology*, *43*(1), 186.

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107*(3), 1084–1092.

Kutscher, L., & Feldman, G. (2019). The impact of past behaviour normality on regret: replication and extension of three experiments of the exceptionality effect. *Cognition and Emotion*, *33*(5), 901–914.

Lagnado, D. A. (2011). 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., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. *Causal learning: Psychology, philosophy, and computa-*

tion, 154–172.

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

Lee, H.-S., Corbera, S., Poltorak, A., Park, K., Assaf, M., Bell, M. D., … others (2018). Measuring theory of mind in schizophrenia research: Cross-cultural validation. *Schizophrenia research*, *201*, 187–195.

Legare, C. H., Sobel, D. M., & Callanan, M. (2017). Causal learning is collaborative: Examining explanation and exploration in social contexts. *Psychonomic bulletin & review*, *24*(5), 1548–1554.

Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology*, *2*(1), 19–32.

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind.". *Psychological review*, *94*(4), 412.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

Levinson, S. C. (2000). *Presumptive meanings*. MIT Press.

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

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lewis, D. (1986). Causal explanation. *Philosophical Papers*, *2*, 214–240.

Lewis, D. (2004). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

Lewis, D. (2013). *Counterfactuals*. John Wiley & Sons.

Liquin, E. G., & Lombrozo, T. (2020, Jun). A functional approach to explanation-seeking curiosity. *Cognitive Psychology*, *119*, 101276. Retrieved from `http://dx.doi.org/10.1016/j.cogpsych.2020.101276` doi: 10.1016/j.cogpsych.2020.101276

Lombard, M., & Gärdenfors, P. (2021). Causal cognition and theory of mind in evolutionary cognitive archaeology. *Biological Theory*, 1–19.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*(2), 273–286.

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

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford: Oxford University Press.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *Oxford handbook of causal reasoning*, 415–432.

Lowe, E. (2001). Event causation and agent causation. *Grazer Philosophische Studien*, *61*(1), 1–20.

Lucas, C. G., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). A rational model of preference learning and choice prediction by children. In *Advances in neural information processing systems* (pp. 985–992).

Ma, L., & Xu, F. (2013). Preverbal infants infer intentional agents from the perception of regularity. *Developmental psychology*, *49*(7), 1330.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Clarendon Press.

Mackie, P. (1992). Causing, delaying, and hastening: Do rains cause fires? *Mind*, *101*(403), 483–500.

Macrae, C. N. (1992). A tale of two curries: Counterfactual thinking and accident-related judgments. *Personality and Social Psychology Bulletin*, *18*(1), 84–87.

Maier, J. (2015). The agentive modalities. *Philosophy and Phenomenological Research*, *90*(1), 113–134.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework.

Personality and Social Psychology Review, *3*(1), 23–48.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014a). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014b, Apr). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186. Retrieved from `http://dx.doi.org/10.1080/1047840x.2014.877340` doi: 10.1080/1047840x.2014.877340

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: the tension between folk concepts and legal concepts of intentionality. *Behav. Sci. Law*, *21*(5), 563–580. Retrieved from `http://dx.doi.org/10.1002/bsl.554` doi: 10.1002/bsl.554

Mandel, D. R. (2005). Counterfactual and causal explanation: From early theoretical views to new frontiers. In D. Mandel, D. Hilton, & P. Catellani (Eds.), *The psychology of counterfactual thinking* (pp. 11–27). New York: Routledge.

Mandel, D. R., & Dhami, M. K. (2005). "what i did" versus "what i might have done": Effect of factual versus counterfactual thinking on blame, guilt, and shame in prisoners. *Journal of Experimental Social Psychology*, *41*(6), 627–635.

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General*, *127*(3), 269–258.

Mandelkern, M., Schultheis, G., & Boylan, D. (2017). Agentive modals. *The Philosophical Review*, *126*(3), 301–343.

Margoni, F., & Surian, L. (2016). Explaining the u-shaped development of intent-based moral judgments. *Frontiers in psychology*, *7*, 219.

Margoni, F., & Surian, L. (2021). Judging accidental harm: Due care and foreseeability of side effects. *Current Psychology*, 1–10.

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.

Maselli, M. D., & Altrocchi, J. (1969). Attribution of intent. *Psychological Bulletin*, *71*(6), 445.

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.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of experimental child psychology*, *141*, 1–22.

McDermott, M. (2002). Causation: Influence versus sufficiency. *The Journal of philosophy*, *99*(2), 84–101.

McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of Personality and Social Psychology*, *79*(5), 677-689. Retrieved from `http://dx.doi.org/10.1037/0022-3514.79.5.677` doi: 10.1037/0022-3514.79.5.677

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, *123*(1-2), 125–148.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: long-term memory for novel acts and multiple stimuli. *Developmental psychology*, *24*(4), 470.

Meltzoff, A. N., Gopnik, A., & Repacholi, B. M. (1999). Toddlers' understanding of intentions, desires and emotions: Explorations of the dark ages.

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-old infants. *Developmental psychology*, *48*(5), 1215.

Mill, J. S. (1875). *A system of logic: Ratiocinative and inductive*. Longmans, Green, Reader, and Dyer London.

Miller, D. T., & McFarland, C. (1986). Counterfactual thinking and victim compensation: A test of norm theory. *Personality and Social Psychology Bulletin*, *12*(4), 513–519.

Miller, D. T., Turnbull, W., & McFarland, C. (1990). Counterfactual thinking and social perception: Thinking about what might have been. In *Advances in experimental social psychology* (Vol. 23, pp. 305–331). Elsevier.

Miller, S. (2018). Joint epistemic action: some applications. *Journal of Applied Philosophy*, *35*(2), 300–318.

Monroe, A., & Ysidron, D. (2021). Not so motivated after all? three replication attempts and a theoretical challenge to a morally motivated belief in free will. *Journal of Experimental psychology. General*, *150(1):e1-e12*. Retrieved from `https:doi://10.1037/xge0000788`

Moore, M. S. (1999). Causation and responsibility. *Social Philosophy and Policy*, *16*(2), 1–51.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PsyArXiv*. Retrieved from `https://psyarxiv.com/upv8t`

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. *PsyArXiv*. Retrieved from `https://psyarxiv.com/nq53z`

Msetfi, R. M., Wade, C., & Murphy, R. A. (2013). Context and time in causal learning: Contingency and mood dependent effects. *PLoS One*, *8*(5), e64063.

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive psychology*, *61*(2), 63–86.

Muentener, P., & Lakusta, L. (2011). The intention-to-cause bias: Evidence from children's causal language. *Cognition*, *119*(3), 341–355.

Murphy, P. K., & Mason, L. (2006). Changing knowledge and beliefs.

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive science*, *41*(2), 447–481.

Murray, S., & Vargas, M. (2018). Vigilance and control. *Philosophical Studies*, 1–19.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, *4*(2), 133–142.

Nelson-le Gall, S. A. (1985). Motive–outcome matching and outcome foreseeability: Effects on attribution of intentionality and moral judgments. *Developmental Psychology*, *21*(2), 332.

Newman, G. E., Keil, F. C., Kuhlmeier, V. A., & Wynn, K. (2010). Early understandings of the link between agents and order. *Proceedings of the National Academy of Sciences*, *107*(40), 17140–17145.

Pan, J. (2016, April 26). 'color play asia' organizer found guilty. *Taipei Times*. Retrieved from `https://www.bbc.co.uk/news/world-asia-33300970`

Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, *121*(1-2), 93–149.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2019, Feb). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. Retrieved from `http://dx.doi.org/10.1145/3241036` doi: 10.1145/3241036

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st ed.). USA: Basic Books, Inc.

Pereboom, D. (2004). Is our conception of agent-causation coherent? *Philosophical Topics*, *32*(1/2), 275–286.

Perugini, M., & Bagozzi, R. P. (2001). The role of desires and anticipated emotions in goal-directed behaviours: Broadening and deepening the theory of planned behaviour. *British Journal of Social Psychology*, *40*(1), 79–98.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & language*, *24*(5), 586–604.

Phillips, J., & Cushman, F. (2017, apr). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*(18), 4649–4654. Retrieved from `https://doi.org/10`

.1073%2Fpnas.1619717114 doi: 10.1073/pnas.1619717114

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General*, *146*(2), 165.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, 65–94. Retrieved from http://dx.doi.org/10.1111/mila.12165 doi: 10.1111/mila.12165

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

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026–1040.

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

Phillips, J., Young, L., & Gerstenberg, T. (n.d.). Normal causation. *unpublished manuscript*.

Piaget, J. (1965). *The moral judgment of the child.(translated by marjorie gabain)*. Routledge & K. Paul (1965, 1932).

Potochnik, A. (2016, Dec). Scientific explanation: Putting communication first. *Philosophy of Science*, *83*(5), 721–732. Retrieved from http://dx.doi.org/10.1086/687858 doi: 10.1086/687858

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112.

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science*, *33*(3), 401–448.

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019, April). Machine behaviour. *Nature*, *568*(7753), 477–486. doi: 10.1038/s41586-019-1138-y

Randall, F., Corcoran, R., Day, J., & Bentall, R. (2003). Attention, theory of mind, and causal attributions in people with persecutory delusions: A preliminary investigation. *Cognitive neuropsychiatry*, *8*(4), 287–294.

"Reddit Traders in r/wallstreet Shake Up the Stock Market". (2016, February 26). *Bloomberg.com.* Retrieved from `https://www.bloomberg.com/news/articles/2020-02-26/reddit-s-profane-greedy-traders-are-shaking-up-the-stock-market`

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

Rivis, A., & Sheeran, P. (2003). Descriptive norms as an additional predictor in the theory of planned behaviour: A meta-analysis. *Current psychology*, *22*(3), 218–233.

Roese, N. J. (1997). Counterfactual thinking. *Psychological bulletin*, *121*(1), 133.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in experimental social psychology* (Vol. 56, pp. 1–79). Elsevier.

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology*, *32*(3), 197–227.

Roese, N. J., & Olson, J. M. (2014). *What might have been: The social psychology of counterfactual thinking*. Psychology Press.

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical perspectives*, *18*, 295–313.

Rosenschein, S. J. (1985). Formal theories of knowledge in ai and robotics. *New generation computing*, *3*(4), 345–357.

Ross, W. D., et al. (1924). *Aristotle's metaphysics* (Vol. 2). Clarendon Press.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive science*, *36*(5), 919–932.

Ryle, G. (2009). *The concept of mind*. Routledge.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of chinese and us preschoolers. *Psychological science*, *17*(1), 74–81.

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

Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*(2), 297–312.

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016a). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, *145*(2), 125–130. Retrieved from `https://doi.org/10.1037%2Fxge0000138` doi: 10.1037/xge0000138

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016b). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, *145*(2), 125.

Samland, J., & Waldmann, M. R. (2014). Do social norms influence causal inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2015). Highlighting the causal meaning of causal test questions in contexts of norm violations. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2092–2097). Austin, TX: Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176. Retrieved from `https://doi.org/10.1016%2Fj.cognition.2016.07.007` doi: 10.1016/j.cognition.2016.07.007

Sartorio, C. (2009). Omissions and causalism. *Noûs*, *43*(3), 513–530.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a bayesian model of theory of mind. *Current opinion in Psychology*, *17*, 15–21.

Saxe, R., Tenenbaum, J., & Carey, S. (2005, dec). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001. Retrieved from `http://dx.doi.org/10.1111/j.1467-9280.2005.01649.x` doi: 10.1111/j.1467-9280.2005.01649.x

Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental psychology*, *43*(1), 149.

Sayre, F. B. (1932). Mens rea. *Harvard Law Review*, *45*(6), 974–1026.

Schaffer, J. (2000, apr). Trumping preemption. *The Journal of Philosophy*, *97*(4), 165. Retrieved from `http://dx.doi.org/10.2307/2678388` doi: 10.2307/2678388

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, *114*(3), 327–358.

Schuster, S., & Degen, J. (2019). Speaker-specific adaptation to variable use of uncertainty expressions. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 7). Cognitive Science Society.

Schwenkler, J., & Sievers, E. (n.d.). Cause," cause", and norm.

Searle, J. R. (1980). The intentionality of intention and action. *Cognitive science*, *4*(1), 47–70.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shakespeare, W. (1858). *Romeo en julia*. AC Kruseman.

Shanks, D. R. (1989). Selectional processes in causality judgment. *Memory & Cognition*, *17*(1), 27–34.

Shaver, K. G., & Drown, D. (1986). On causality, responsibility, and self-blame: A theoretical note. *Journal of personality and social psychology*, *50*(4), 697.

Sher, G. (2009). Who knew. *Oxford University Press USA. Smart, JJC (1961).'Free-Will, Praise and Blame'. Mind*, *70*, 291–306.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*(1), 1–51.

Shvo, M., Klassen, T. Q., & McIlraith, S. A. (2020). Towards the role of theory of mind in explanation. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 75–93).

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=afex` (R package version 0.26-0)

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press, USA.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, *66*, 223–247.

Slugoski, B. R., Lalljee, M., Lamb, R., & Ginsburg, G. P. (1993). Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, *23*(3), 219–238.

Smith, H. (1983). Culpable ignorance. *The Philosophical Review*, *92*(4), 543–571.

Sobel, D., & Sommerville, J. (2010). The importance of discovery in children's causal learning from interventions. *Frontiers in Psychology*, *1*, 176.

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779.

Spelke, E. S. (1990, jan). Principles of object perception. *Cognitive Sci-*

ence, *14*(1), 29–56. Retrieved from `http://dx.doi.org/10.1207/s15516709cog1401_3` doi: 10.1207/s15516709cog1401_3

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

Spellman, B. A., & Gilbert, E. A. (2014). Blame, cause, and counterfactuals: The inextricable link. *Psychological Inquiry*, *25*(2), 245–250.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language*, *17*(1-2), 3–23. doi: 10.1111/1468-0017.00186

Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in inaction: Counterfactual simulation and causation by omission. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1132–1137). Austin, TX: Cognitive Science Society.

Sterelny, K. (1990). *The representational theory of mind: An introduction.* Basil Blackwell.

Strevens, M. (2008). *Depth.* Harvard University Press.

Strickland, B., Silver, I., & Keil, F. C. (2017, April). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory & Cognition*, *45*(3), 442-455. doi: 10.3758/s13421-016-0668-x

Sytsma, J. (2019a). The character of causation: Investigating the impact of character, knowledge, and desire on causal attributions.

Sytsma, J. (2019b). The effects of single versus joint evaluations on causal attributions.

Sytsma, J. (2020a). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, 1–21.

Sytsma, J. (2020b). Resituating the influence of relevant alternatives on attributions.

Sytsma, J. (2021). The responsibility account.

Sytsma, J., & Livengood, J. (2019). Causal attributions and the trolley problem.

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 Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.

"Taiwan Formosa Water Park explosion injures hundreds". (2015, June 28). *BBC World Asia*. Retrieved from `https://www.bbc.co.uk/news/world-asia-33300970`

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive science*, *12*(1), 49–100.

Turley, K. J., Sanna, L. J., & Reiter, R. L. (1995). Counterfactual thinking and perceptions of rape. *Basic and Applied Social Psychology*, *17*(3), 285–303.

Turnbull, W., & Slugoski, B. R. (1988). Conversational and linguistic processes in causal attribution. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 66–93). Brighton, UK: Harvester Press.

Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: Heuristics and biases*. Springer.

U.S. Chemical Safety and Hazard Investigation Board. (2006). Investigation report. combustible dust hazard study. Retrieved from `https://www.csb.gov/combustible-dust-hazard-investigation`

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100.

Vallée-Tourangeau, F., Murphy, R. A., & Baker, A. (2005). Contiguity and the outcome density bias in action–outcome contingency judgements. *The Quarterly Journal of Experimental Psychology Section B*, *58*(2b), 177–192.

van Fraassen, B. (1980). *The scientific image*. Oxford University Press.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018, April). Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*. doi: 10.1111/cogs.12605

Waismeyer, A., Meltzoff, A. N., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month-olds: an action measure. *Developmental science*, *18*(1), 175–182.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*(6), 307–311.

Walsh, C. R., & Byrne, R. M. (2007b). How people think "if only..." about reasons for actions. *Thinking & Reasoning*, *13*(4), 461–483.

Walsh, C. R., & Byrne, R. M. J. (2007a, Oct). How people think "if only ..." about reasons for actions. *Thinking & Reasoning*, *13*(4), 461-483. Retrieved from `http://dx.doi.org/10.1080/13546780701382120` doi: 10.1080/13546780701382120

Waskan, J., Harmon, I., Horne, Z., Spino, J., & Clevenger, J. (2014). Explanatory anti-psychologism overturned by lay and scientific case classifications. *Synthese*, *191*(5), 1013–1035.

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? a meta-analysis of the experimental evidence. *Psychological bulletin*, *132*(2), 249.

Weinstein, A. G. (1972). Predicting behavior from attitudes. *Public Opinion Quarterly*, *36*(3), 355–360.

Wellman, H. M. (2011). Developing a theory of mind.

Wellman, H. M., & Banerjee, M. (1991). Mind and emotion: Children's understanding of the emotional consequences of beliefs and desires. *British Journal of Developmental Psychology*, *9*(2), 191–214.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of personality and social psychology*, *56*(2), 161.

Westra, E., & Carruthers, P. (2017). Pragmatic development explains the theory-of-mind scale. *Cognition*, *158*, 165–176.

Whiten, A. (2002). Imitation of sequential and hierarchical structure in action: Experimental studies with children and chimpanzees.

Widerker, D. (2017). *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Routledge.

Wieland, J. W., & Robichaud, P. (2017). *Responsibility: The epistemic condition*. Oxford University Press.

Wiener, R. L., & Pritchard, C. C. (1994). Negligence law and mental mutation. In *Applications of heuristics and biases to social issues* (pp. 117–136). Springer.

Wilkenfeld, D. A., & Lombrozo, T. (2018). Explanation classification depends on understanding: extending the epistemic side-effect effect. *Synthese*, 1–28.

Willemsen, P. (2016). Omissions and expectations: A new approach to the things we failed to do. *Synthese*. Advance online publication. doi: 10.1007/s11229-016-1284-9

Willemsen, P. (2018). Omissions and expectations: A new approach to the things we failed to do. *Synthese*, *195*(4), 1587–1614.

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), e12562.

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

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

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.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. In *Psychology of learning and motivation* (pp. 167–202). Elsevier BV. Retrieved from `http://dx.doi.org/10.1016/b978-0-12-407237`

-4.00005-0 doi: 10.1016/b978-0-12-407237-4.00005-0

Woo, B. M., Steckler, C. M., Le, D. T., & Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition*, *168*, 154–163.

Woodward, J. (2001). Causation and manipulability.

Woodward, J. (2002). What is a mechanism? a counterfactual account. *Philosophy of Science*, *69*(S3), S366–S377.

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

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

Woodward, J. (2007). Interventionist theories of causation in psychological perspective. *Causal learning: Psychology, philosophy, and computation*, 19–36.

Woodward, J. (2011). Psychological studies of causal and counterfactual reasoning. *Understanding counterfactuals, understanding causation. Issues in philosophy and psychology*, 16–53.

Woodward, J. (2014). *A functional account of causation.* Retrieved from http://philsci-archive.pitt.edu/10978/

Wright, L. (1976). *Teleological explanations. an etiological analysis of goals and functions*. University of California Press.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2021). A survey of human-in-the-loop for machine learning. *arXiv preprint arXiv:2108.00941*.

Wu, Y., Muentener, P., & Schulz, L. E. (2016). The invisible hand: toddlers connect probabilistic events with agentive causes. *Cognitive science*, *40*(8), 1854–1876.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*(2), 245–280.

Yee, T. W., et al. (2010). The vgam package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016, April). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and*

Language, *87*, 128–143. doi: 10.1016/j.jml.2015.08.003

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, *119*(2), 166–178.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *neuroimage*, *40*(4), 1912–1920.

Young, L., & Saxe, R. (2009). An fmri investigation of spontaneous mental state inference for moral judgment. *Journal of cognitive neuroscience*, *21*(7), 1396–1405.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214.

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and personality psychology compass*, *7*(8), 585–604.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, *107*(3), 410–426.

Zufferey, S. (2010). Lexical pragmatics and theory of mind. *The Acquisition of*.