Changing Minds — Epistemic Interventions in Causal Reasoning

Lara Kirfel*
University College London

David Lagnado
University College London

Abstract

Did Tom's use of nuts in the dish cause Billy's allergic reaction? According to counterfactual theories of causation, an agent is judged a cause to the extent that their action made a difference to the outcome (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2020; Gerstenberg, Halpern, & Tenenbaum, 2015; Halpern, 2016; Hitchcock & Knobe, 2009). In this paper, we argue for the integration of epistemic states into current counterfactual accounts of causation. In the case of ignorant causal agents, we demonstrate 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 the agent's actual causal action. In four experiments, we show that people's causal judgment as well as their reasoning about alternatives is sensitive to the epistemic conditions of a causal agent: Knowledge vs. ignorance (Experiment 1), self-caused vs. externally caused ignorance (Experiment 2), the number of epistemic actions (Experiment 3), and the epistemic context (Experiment 4). We see two arguments for integrating epistemic states into causal models and counterfactual frameworks. First, assuming the intervention on indirect, epistemic causes might allow us to explain why people attribute decreased causality to ignorant vs. knowing causal agents. Moreover, causal agents' epistemic states pick out those factors that can be controlled or manipulated in order to achieve desirable future outcomes, reflecting the forward-looking dimension of causality. We discuss our findings in the broader context of moral and causal cognition.

Keywords: epistemic states, knowledge, counterfactuals, causal judgment, moral judgment

*Corresponding author: Lara Kirfel (ucjulki@ucl.ac.uk), 26 Bedford Way, London WC1H 0AP

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")

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 reached him. The messenger instructed to deliver the letter is prevented by an outbreak of the plague. Not knowing that Juliet is alive — and that his own death would later cause Juliet to stab herself out of grief — Romeo poisons himself. Before he drinks the poison, Romeo is however 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 tragic 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.

Aim of this paper

What is the most relevant change in an alternative ending of "Romeo and Juliet"? Intuitively, it is not (just) a change of what Romeo did, but rather, what he knew, or more specifically, *didn't* know. While causal models can in principle include both physical as well as mental variables (Sloman, Fernbach, & Ewing, 2009, 2012), our aim in this paper is to suggest an extended focus of counterfactual theories 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, i.e. the action that caused the outcome. Integrating epistemic states into causal models and counterfactual frameworks allows us to explain why people often attribute decreased causality to ignorant agents (Hilton, McClure, & Moir, 2016; Hilton & Slugoski, 1986; Kirfel & Lagnado, 2021; Lombrozo, 2010; Samland & Waldmann, 2016a). We present an extension to current counterfactual accounts

by introducing epistemic state variables, and we test this extension by investigating people's causal judgments in four experiments.

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 attribute 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). Responsibility judgments decrease to the extent that the caused harm was unintended or unforeseen (Cushman, 2015; Engelmann & Waldmann, 2021; Margoni & Surian, 2021; Nelson-le Gall, 1985; Young & Saxe, 2008). Theoretical models of moral judgements assume agentive epistemic states to be an integral and early criterion in the process of moral decision formation (Alicke & Rose, 2012; 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).

Causation by 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, 2020; Lagnado & Channon, 2008; Lombrozo, 2010). Agents lacking knowledge (Gilbert, Tenney, Holland, & Spellman, 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, Hilton, & Sutton, 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, people's preference for abnormal actions as causes has been shown to be moderated by the agents' knowledge states about their actions (Kirfel & Lagnado, 2020; Kirfel & Phillips, 2021; Samland, Josephs, Waldmann, & Rakoczy, 2016). There is a close connection between the development of causal reasoning and theory of mind: 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 discussed in the previous section, 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, Rose, & Bloom, 2012; Samland & Waldmann, 2016a; Sytsma, 2019), 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 assess the genuine causality of agents. One account proposes that knowing and intentional agents are perceived as more "robust" causes than ignorant and unintentional ones (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Lombrozo, 2010; Murray & Lombrozo, 2017; Phillips & Shaw, 2015). At the core of "counterfactual robustness" (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 knowledgeable 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 (a car accident because of malfunctioning brakes), participants generate more counterfactuals about ways the bad outcome could have been prevented that the actor could control (e.g. informing the driver about the issue, fixing the brakes) (Spellman & Gilbert, 2014). Knowledge affects causation ratings by increasing the generation of controllable ways the outcome could have been different.

In sum, *if* the influence of agent epistemic states on causal judgments reveals something about people's causal thinking, first evidence suggests that this influence might work via people's reasoning about alternatives. 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 the general 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.

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; Mackie, 1974). While this framework has been originally developed as a

normative theory of causation (Lewis, 2013), it has been adapted to address people's causal intuitions (Halpern & Hitchcock, 2015) and extended to capture people's fine-grained 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), removing the causal candidate from the scene or even perturbing certain properties (e.g. position, velocity) of the cause (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, Paul, et al., 2003), or *overdetermination* (Gerstenberg, Halpern, & Tenenbaum, 2015; Lagnado, Gerstenberg, & Zultan, 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, Lagnado, & Tenenbaum, 2015; Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Icard, 2020; Henne, Kulesza, Perez, & Houcek, 2021; Icard, Kominsky, & Knobe, 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 problems. 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: 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; Quillien & German, 2021) and responsibility or blame (Chockler & Halpern, 2004; Gerstenberg, Halpern, & Tenenbaum, 2015), the general framework in its current form does not quite capture the influence of epistemic states discussed earlier (Hilton et al., 2016) — the perceived causal difference between a causal agent who is ignorant vs. knowledgeable. In this paper, we argue for counterfactual theories to extend their focus for mental states.

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

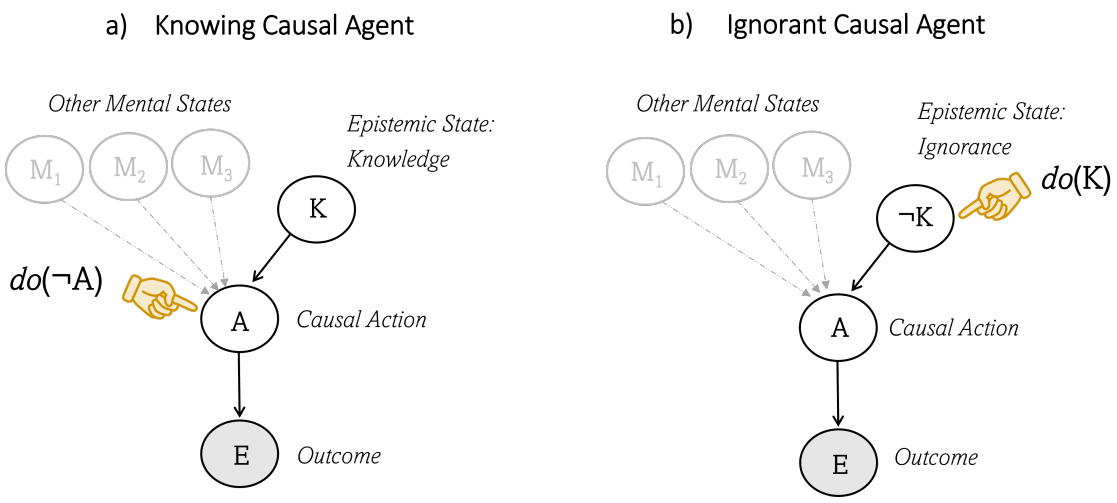

Figure 1. **Causal Models including Epistemic States.** If an agent knows about the causal outcome of their action (1a), counterfactual intervention targets the agent's action A. In case of ignorant causal agents, counterfactual intervention targets the agent's epistemic state of ignorance $\neg K$ (1b).

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 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 or Quillien and German (2021) for a causal definition of 'intentional' action). 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. To account for the causality of mental states, interventionist causal theories assume the intervention on psychological variables (Campbell, 2007; Kaiserman, 2020), although these kind of interventions have often been argued to represent "soft" interventions (Campbell,

2007; Eronen, 2020; Kaiserman, 2020).[2] We propose that people intervene on mental states, and here specifically, *epistemic* states, in order to assess the causality of ignorant causal agents as a whole.

What does this mean exactly? Let us consider again a case in which an agent performs a certain action A, e.g. doctor Jones gives her patient a new drug, and either knows K (Figure 1a) or does not know $\neg K$ that this drug has a certain side effect (Figure 1b).[3] 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 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. 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 arguments for this claim.

Why Epistemic Intervention?

Changing Actions via Knowledge. On the one hand, epistemic states, and mental states more broadly, 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). Walsh and Byrne (2007) as well

[2] Can we intervene on an agent's intention to do A without also changing the agent's belief that A is beneficial or their reason for doing A? Campbell (2007) argues that such 'surgical', selective interventions on specific mental states are often counter-intuitive, violating the assumptions of agent rationality. According to a 'mental holism' view, mental states are connected in a web of mental states, such that it is not possible to intervene on one selectively while holding fixed others. Allowing for 'soft interventions' that also target causal variables such as beliefs or reasons that go into an 'intention variable' might account for how intervention on psychological states work (Campbell, 2007, 2010), while still indicating the causality of the intervened variable (Eberhardt & Scheines, 2007). In this study, we focus on epistemic states about specific propositions about the world, which we tentatively assume to be more modular mental states. We are open to the idea that epistemic interventions represent soft interventions, and that for a knowledge state to change, an intervention on prior causal variables such as prior beliefs etc. is necessary. We explicitly allow for other mental states to be causally *affected* by epistemic interventions, and this assumption is part of our central argument.

[3] What does mean for an agent to be ignorant of a proposition φ? According to the 'Standard View' in epistemology, ignorance is defined as *not knowing* and can be formalised as $\neg K\varphi$. Le Morvan (2013) further distinguishes between *propositional ignorance*, not knowing (of) φ, from *factive ignorance*, not knowing *that* φ. Factive ignorance includes ignorance of the truth-conditions associated with a proposition: the agent doesn't know whether φ is true or not, but is able to consider or entertain that φ. Propositional ignorance in contrast precludes being able to even consider or entertain that φ ("Socrates is ignorant that TikTok is the most downloaded app in 2020"; $\neg K\varphi \wedge \neg K\neg\varphi$) (Meyer & Hoek, 1995; Van Der Hoek & Lomuscio, 2004). In this paper, the epistemic state of interest will be a state of factive ignorance. Newer accounts argue that ignorance is a lack of true belief (Peels, 2010, 2012), formalised by a special epistemic operator $I\varphi$ (Kubyshkina & Petrolo, 2019; Van Der Hoek & Lomuscio, 2004)

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 alternative scenarios (see also Bonnefon, 2007; Juhos, Quelhas, & Byrne, 2015). 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 have a reason 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 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 of action given some state of knowledge will depend on further assumptions such as the agent's general preferences, motivations for the action, etc (Weinstein, 1972). These assumptions vary depending on both context as well as people's individual priors. However, assessing the likelihood of a difference in the outcome for in counterfactual reasoning about ignorant agents is 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). Given that the agent's action is a necessary cause for the effect, in contrast to a case in which the intervention directly targets the causal action, intervening on a prior (epistemic) variable naturally 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.

Future Causation. In addition, we think 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 (Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017; 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., 2020; Lombrozo, 2010; Vasilyeva, Blanchard, & Lombrozo, 2018). 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; Ker-

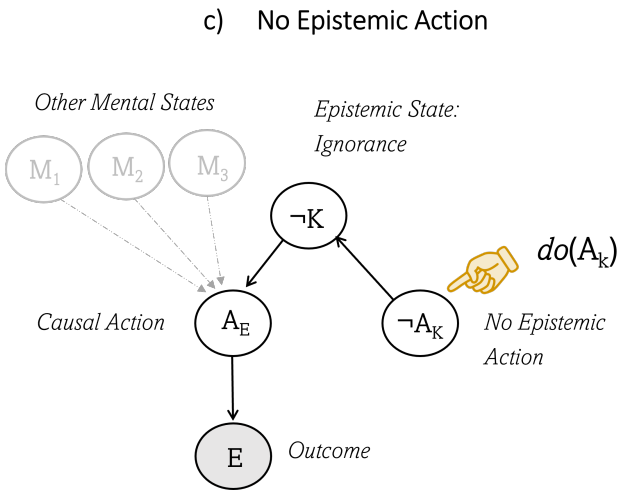

Figure 2. **Causal Models including Epistemic States and Epistemic Action.** When the epistemic conditions for knowledge acquisition are known, counterfactual interventions target the agent's epistemic actions A_{K} that could update their state of knowledge.

wer, Rosman, Wedderhoff, & Chasiotis, 2021; Murphy & Mason, 2006). Only an agent who possesses knowledge about the consequences of their action can or will effectively adapt their behaviour (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 of causality might also explain why people assign less causality to agents bringing about accidental positive outcomes (Guglielmo & Malle, 2019; Lagnado & Channon, 2008; Malle, Guglielmo, & Monroe, 2014). 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 epistemic states hence also marks an intervention that makes the agent a robust cause for the targeted outcome state (Grinfeld et al., 2020).

Epistemic Actions. Once we acknowledge the role of epistemic states in people's causal models, this also brings up 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; Miller, 2018), — actions the agent could have performed in order to gain the relevant knowledge (Figure 2). 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 $\mathrm{A_K}$ that are connected to the epistemic state K, and causal actions $\mathrm{A_E}$ that are connected to the outcome E (Figure 2).

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. Counterfactual theories of causation consider whether counterfactual dependence is obtained in the actual world, but also under different 'contingencies', i.e. when background 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.

Hypotheses

In this paper, 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 Different 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 more causal than an agent who would remain ignorant in both actual and possible worlds.

Individual Causal Models of the World

The central assumption underlying our hypotheses 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 the ending of "Romeo and Juliet", 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? What if he had checked Juliet's pulse? 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. 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.

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). Blameworthiness 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 led them to gain 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, 2016a; 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.

Experiment 1

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

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 $N = 121$ ($M_{\text{age}} = 38.42$, $SD_{\text{age}} = 11.15$, $N_{\text{female}} = 40$). We adopted a 2 knowledge (knowledge vs. no knowledge) $\times$ 3 scenario ("hospital" vs. "garden" vs "bakery") design. 'Knowledge' was manipulated within participants and 'scenario' was manipulated between participants.

Material and Procedure

Participants read both the 'knowledge' as well as the 'no knowledge' condition of one of the three scenarios ("hospital", "garden", "bakery"; see Appendix A) 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, a 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."

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').

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 imagined alternative scenarios.

At the end of the experiment, participants provided demographic information and were thanked for their participation in the study.[4]

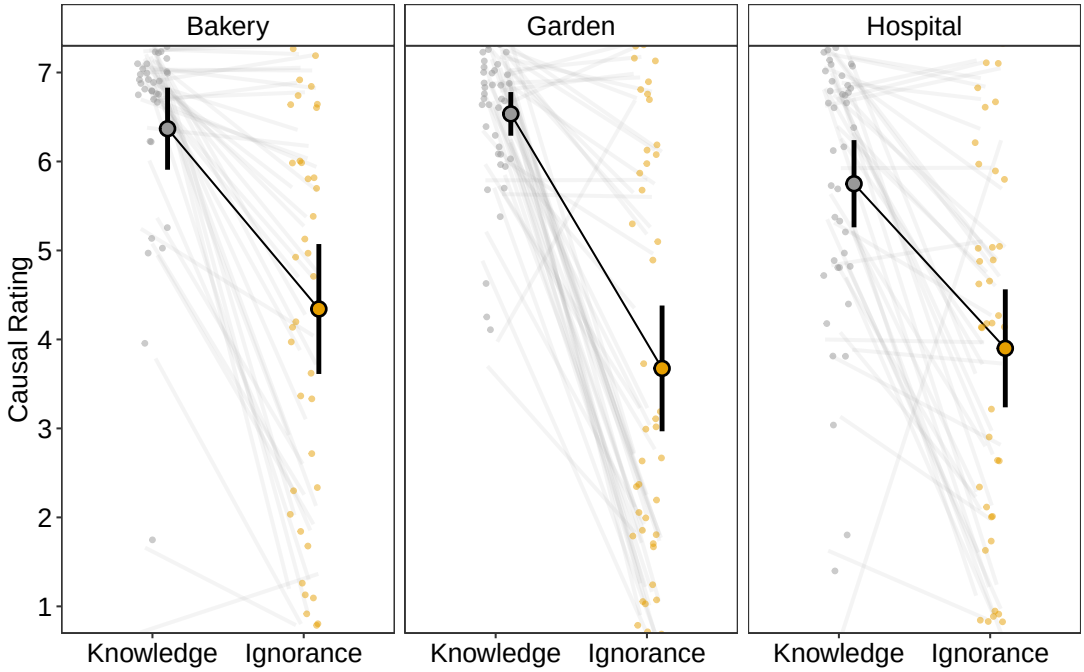

Figure 3. **Experiment 1: Causal Ratings**. Big dots are group means. Error bars depict 95% Confidence Intervals. Small dots connected by lines are individual participants' judgments, jittered for visibility.

Analysis of Data

The central manipulation in our experiment — the knowledge manipulation — was employed as a within-participants variable. To control for repeated measure-effects, we analysed participants responses to the different knowledge manipulations both as *within-subject* as well as *between-subject* contrast in all four experiments.

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. 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

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

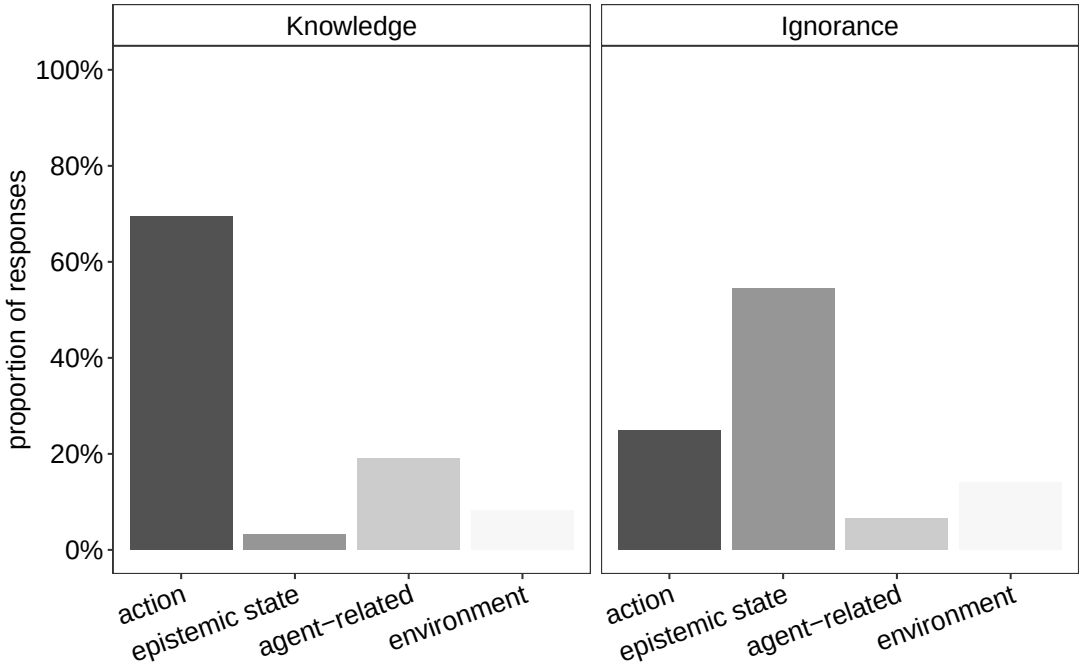

Figure 4. **Experiment 1: Counterfactual responses**. Proportion of choice of four counterfactual response categories ("action", "epistemic state", "agent-related", "environment")

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.

Results

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). 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].

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%).

- **"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..."*).

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 4). 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 4).

Discussion

The first experiment replicated previous findings demonstrating the influence of agent epistemic states on people's causal attributions (Gilbert et al., 2015; Hilton et al., 2016; 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 about the consequences of their action shifts the target of the counterfactual intervention when thinking about how things could have gone differently. In case of an agent's ignorance, people are less likely to refer to a change in the agent's causal action, but prefer to envisage a change in the agent's epistemic state, and more precisely, a change from ignorance to knowledge. 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 very first scenario they saw in the study. Experiment 1 provides evidence for our hypothesis that people naturally represent and refer to agents' 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., 2018; 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 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 are underspecified. In the "ignorance" condition, our experimental scenarios leave open whether and to what extent the agent could have changed acquired the relevant knowledge. Hypothesis 2 predicts that the epistemic conditions of an ignorant causal agent 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. Addressing this question was the aim of Experiment 2.

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 5). 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 5 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 5 b).

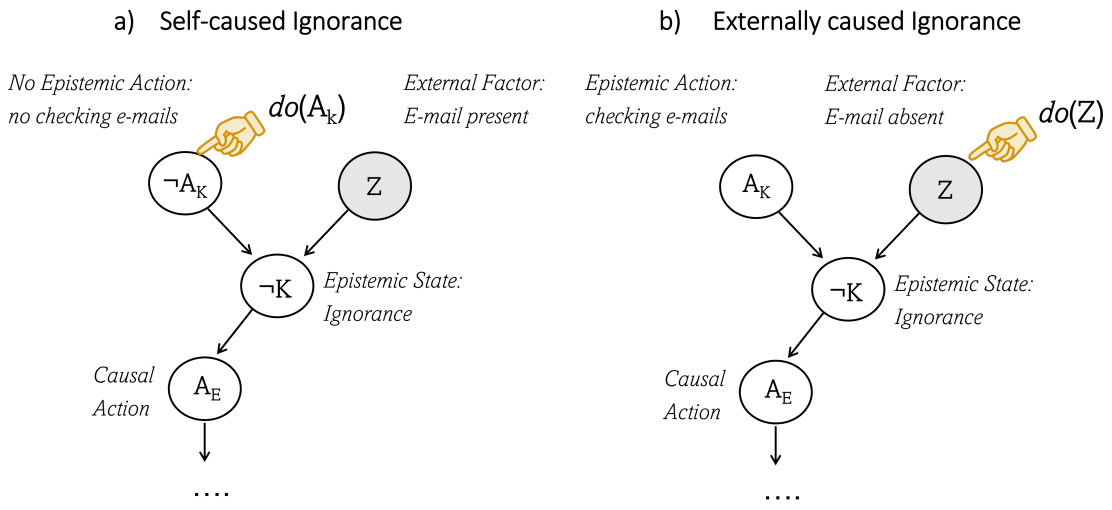

Figure 5. **Experimental Conditions of Experiment 2.** In the "Self-caused Ignorance" condition (a), the agent does not perform the epistemic action that would update their knowledge, although the external condition are given, i.e. the information is available. In the "Externally Caused Ignorance" condition (b), the agent aims to inform themselves, but the information is not available.

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 ($M_{\text{age}} = 37.78$, $SD_{\text{age}} = 11.67$, $N_{\text{female}} = 59$). We adopted a 2 ignorance (self-caused vs. externally caused) $\times$ 3 scenario ("hospital" vs. "garden" vs "bakery") design. 'Ignorance' was manipulated within participants and 'scenario' was manipulated between participants.

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 scenario ends with the agent applying the relevant item, ignorant about the harmful properties of the item. As a result, a bad effect obtains.

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]".

Knowledge and Blame Rating. In addition to causal and counterfactual responses, we wanted to assess people's judgments about the agent's possibilities for knowledge as well as the agent's blameworthiness for their state of 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 own 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').

Results

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 6). There was no significant effect of scenario ($p = .90$) [*between-contrast*: $p = .72$] and no interaction between ignorance and scenario ($p = .99$) [*between-contrast*: $p = .44$].

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

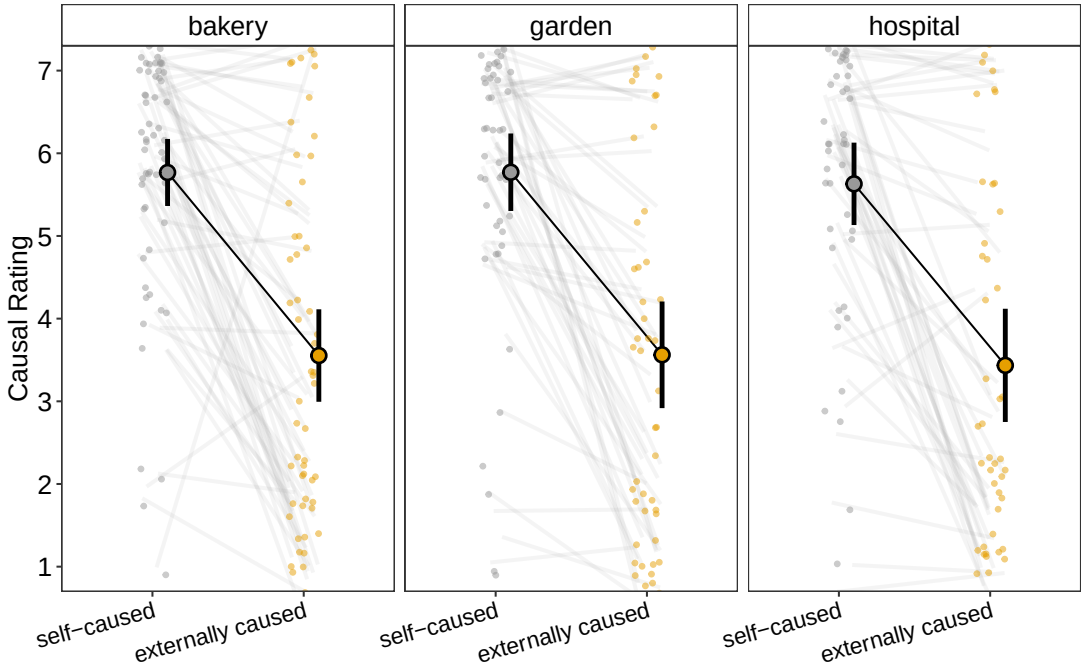

Figure 6. **Experiment 2: Causal Ratings**. Big dots are group means. Error bars depict 95% Confidence Intervals. Small dots connected by lines are individual participants' judgments, jittered for visibility.

- **"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

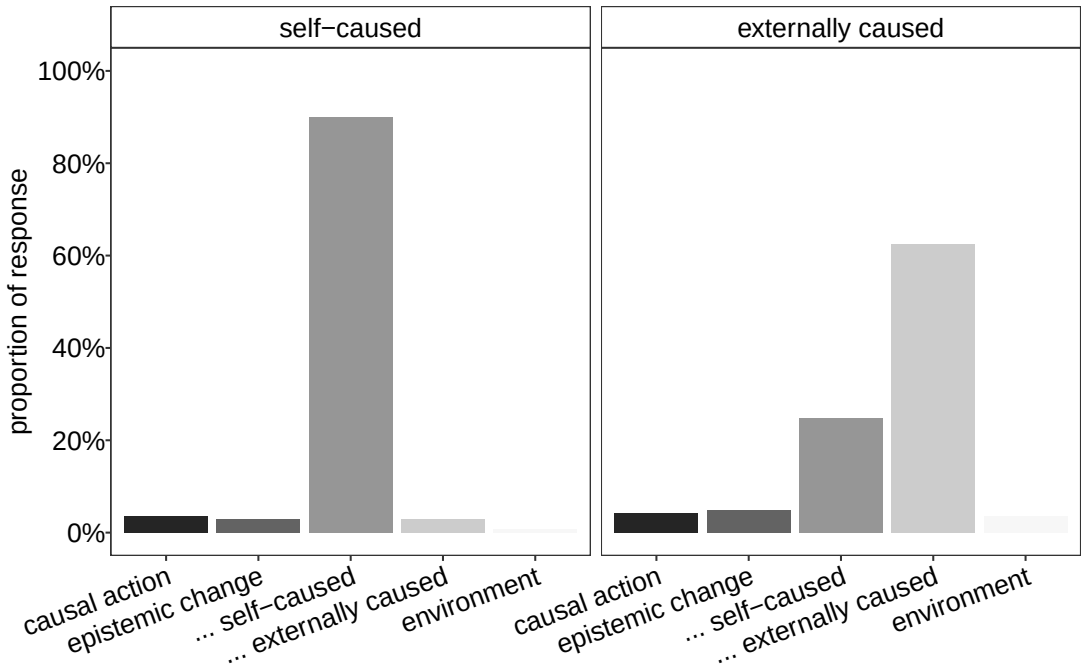

Figure 7. **Experiment 2: Counterfactual responses**. Proportion of responses in the four counterfactual response categories ("action", "epistemic state change", "self-caused epistemic change", "externally caused epistemic change", "epistemic change by other", "environment")

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

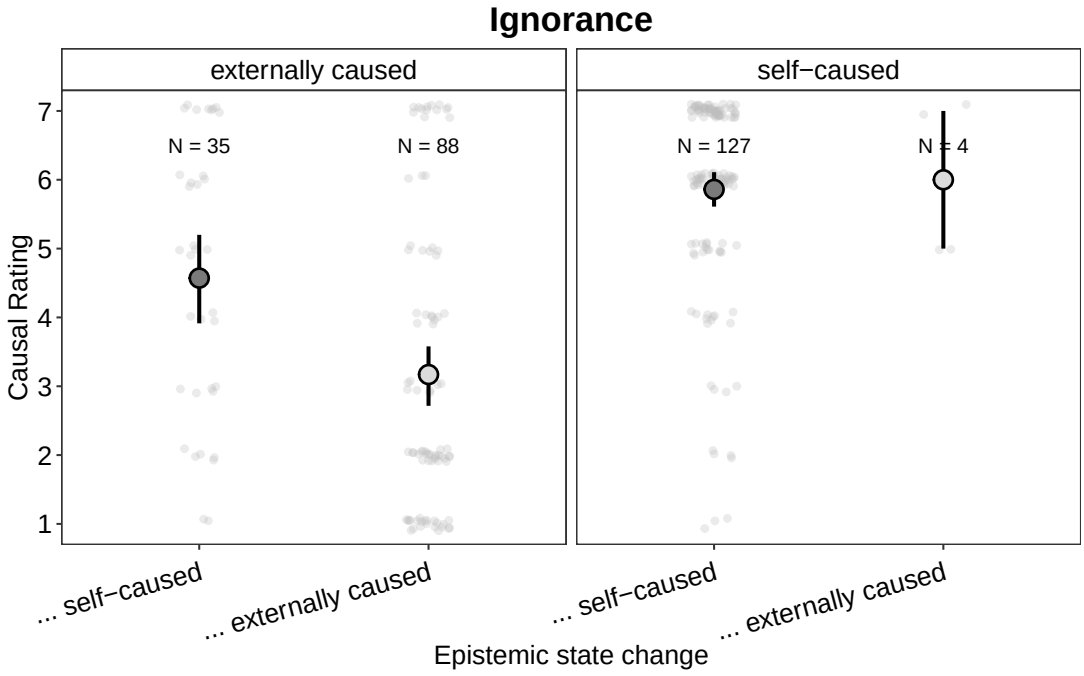

Figure 8. **Experiment 2: Causal Ratings by Counterfactual Response Category**. Causal Ratings of participants who gave imagined the agents' epistemic state change to be self or externally caused, split by ignorance condition

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 7)

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%).

Subgroup Analysis: Causal Rating by Counterfactual Response. In the externally caused ignorance condition, there was a substantial proportion of people who in-

dicated 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 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 9).

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*: $t(121) = -0.17$, $p = .86$].

Knowledge Rating. The condition under which ignorance came about significantly predicted people's judgement about the mutability of the agent's epistemic state, $\chi^2(1) = 114.50$; $p < .001$ [*between-subject contrast*: $F(1) = 56.18$; $p < .001$]. People 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 8).

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 ($b = 0.69$, $SE = .25$, $t = -2.76$, $p = .02$).

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*: $F(1) = 37.58$; $p < .001$].

Discussion

The results of Experiment 2 show that the epistemic conditions of ignorance influence people's causal judgements about an ignorant causal 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. In dependence of whether the access to relevant information is prevented by an external cause or the agent's own actions, people differ in how

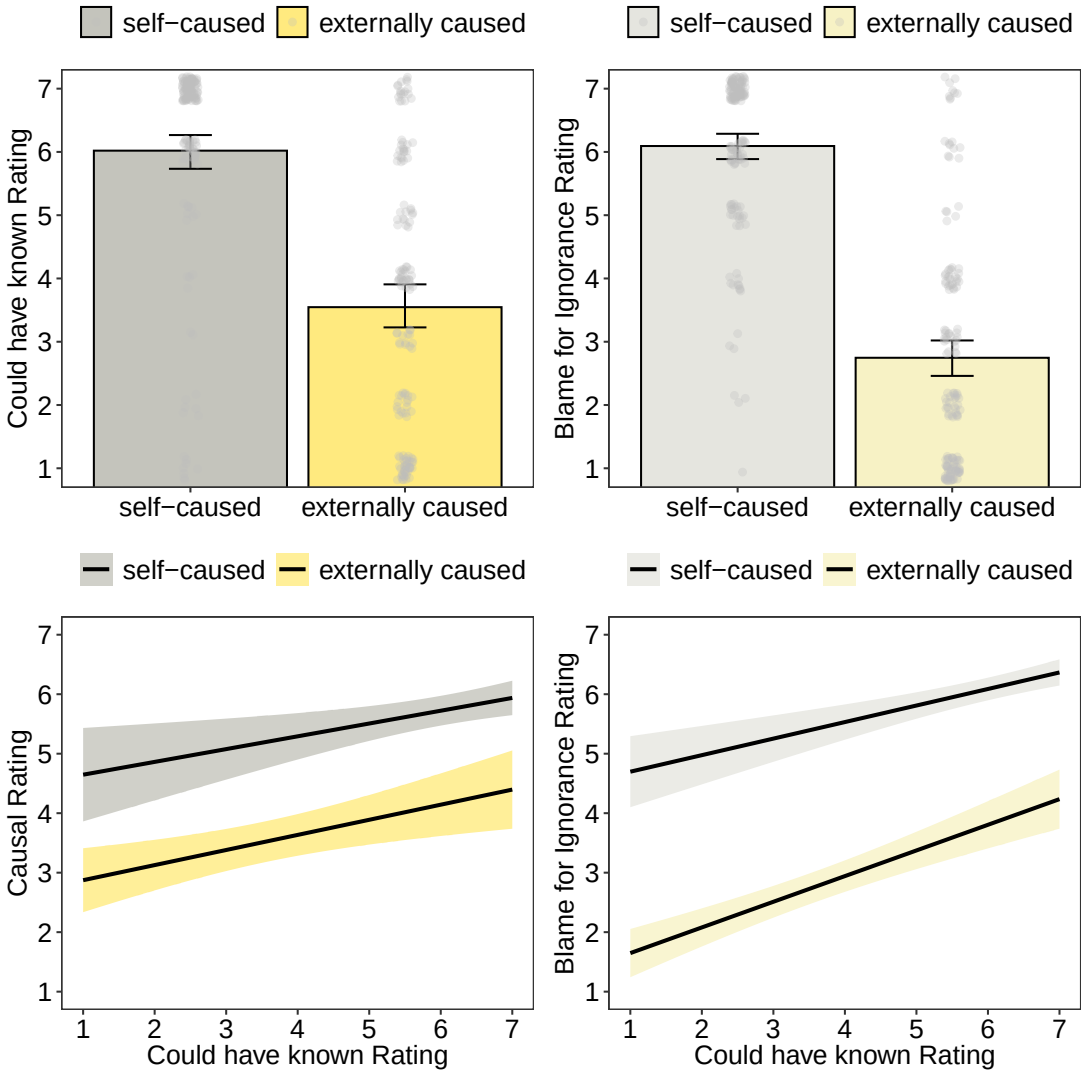

Figure 9. **Experiment 2: Knowledge and Blame Ratings**. Bar graphs depict means of i) "Could have known" ratings and ii) "Blameworthiness for Ignorance" ratings. Regression plots depict "Could have known" ratings as predictor for i) Causal ratings and ii) Blame for Ignorance ratings

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 agent could have performed. 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

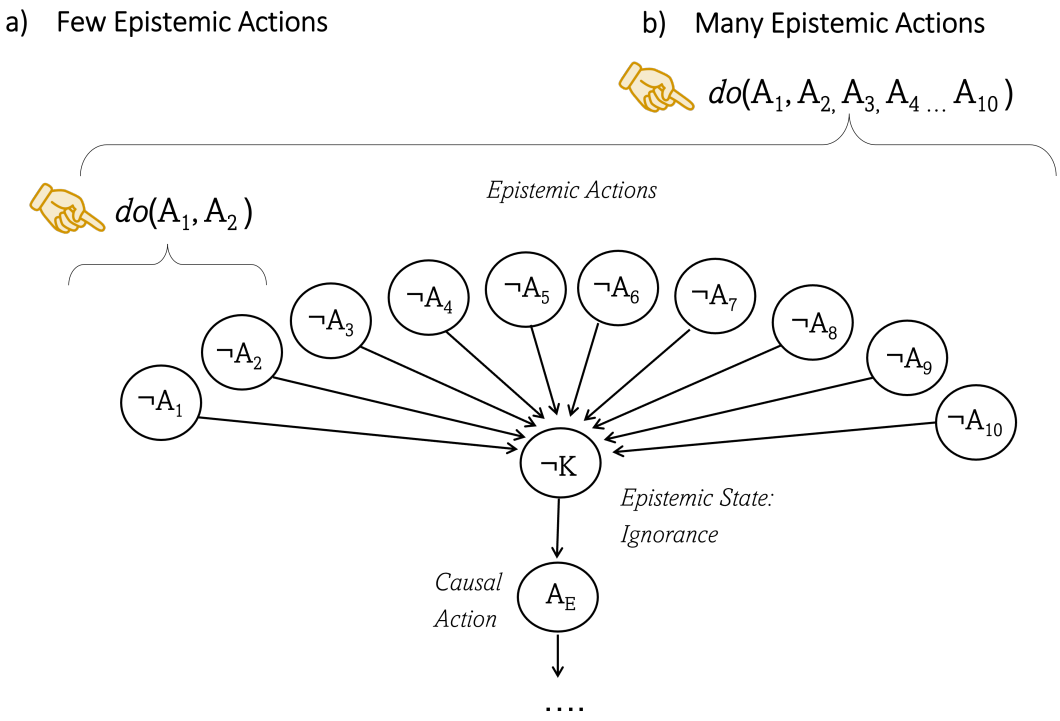

Figure 10. **Experimental Conditions of Experiment 3.** In the "Few Epistemic Actions necessary" condition (a), it takes the agent two separate actions in order access information that would update their knowledge. In the "Many Epistemic Actions necessary" condition (b), the agent needs to perform ten actions of inquiry in order to access the relevant information that would result in knowledge.

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.

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

Table 1
Condition "Many Epistemic actions necessary"

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 2
Condition "Few Epistemic actions necessary"

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**

few epistemic actions (Figure 10a) or many epistemic actions (Figure 10b) in order to acquire knowledge.

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 ($M_{\text{age}} = 35.12$, $SD_{\text{age}} = 11.27$, $N_{\text{female}} = 27$). We adopted a 2 ignorance (few actions vs. many actions) $\times$ 3 scenario ("hospital" vs. "garden" vs "bakery") design. 'Ignorance' was manipulated within participants and 'scenario' was manipulated between participants.

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 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 in line with the sensitivity of the test, nine tests (few epistemic actions condition) or only one of the ten tests (many epistemic 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 1 and Table 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 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 e-mail of a series 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 epistemic actions condition" (only e-mail '10' contains a positive test result, see Table 1), and at least one more e-mail in the "few epistemic actions" condition (e-mails '2'-'10' contain a positive test result, Table 2). As before, the doctor prescribes the drug to a patient with the specific blood type and the patient is harmed.

Results

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 11). 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 ($p = .87$).

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 ($N = 28$) (*"If Alex does not [sic] use the fertilizer splendor the bourbon roses would not have died"*).

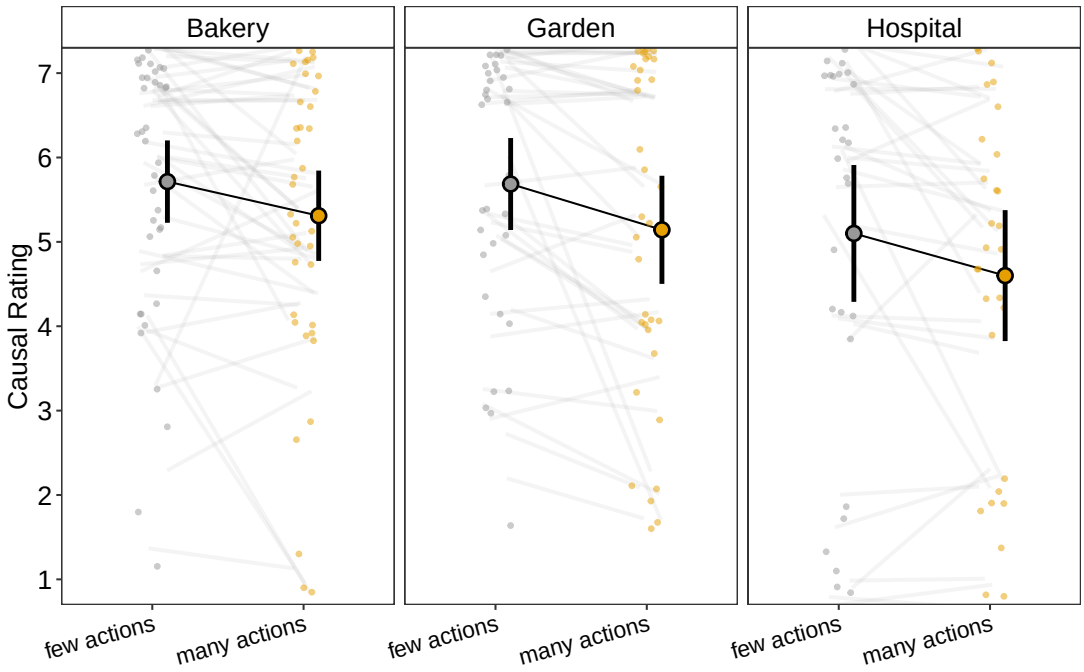

Figure 11. **Experiment 3: Causal Ratings**. Big dots are group means. Error bars depict 95% Confidence Intervals. Small dots connected by lines are individual participants' judgments, jittered for visibility.

- **"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

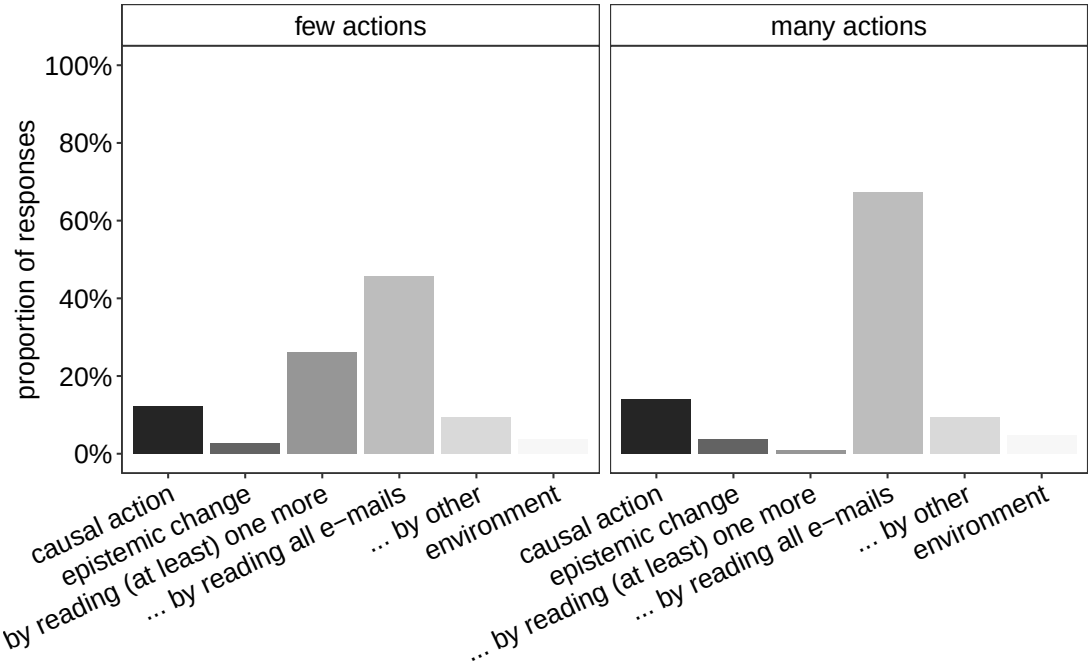

Figure 12. **Experiment 3: Counterfactual responses**. Proportion of responses in the four counterfactual response categories ("action", "epistemic state change", "epistemic state change by reading (at least) one more e-mail", "epistemic state change by reading all e-mails", "'epistemic change by other", "environment")

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%).

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 "counterfactual 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 13).

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

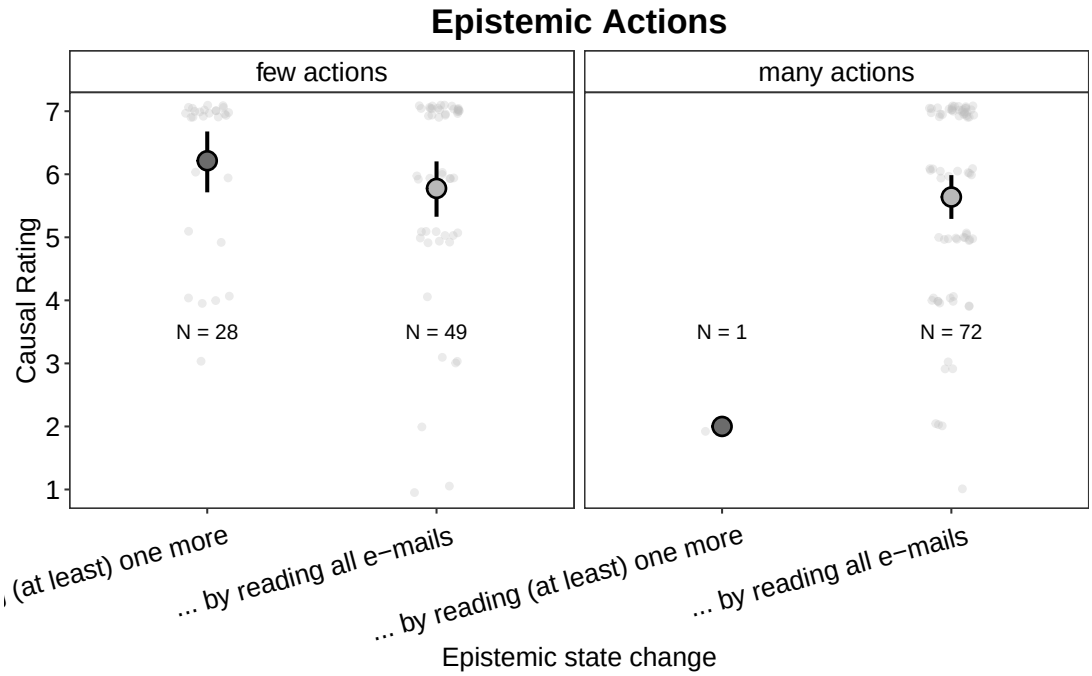

Figure 13. **Experiment 3: Causal Ratings by Counterfactual Response Category**. Causal Ratings of participants who gave imagined the agents' epistemic state change by few vs. many epistemic actions, split by ignorance condition

= 2.09, 95% CI [4.97, 5.21]) compared to fewer 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 epistemic actions" condition M= 5.33, SD = 2.09, 95% CI [4.97, 5.21]) compared the "few epistemic 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 14) [*between-subject contrast*: $F(1) = 53.37$; $p < .001$].

Discussion

The results of Experiment 3 show 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 found 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

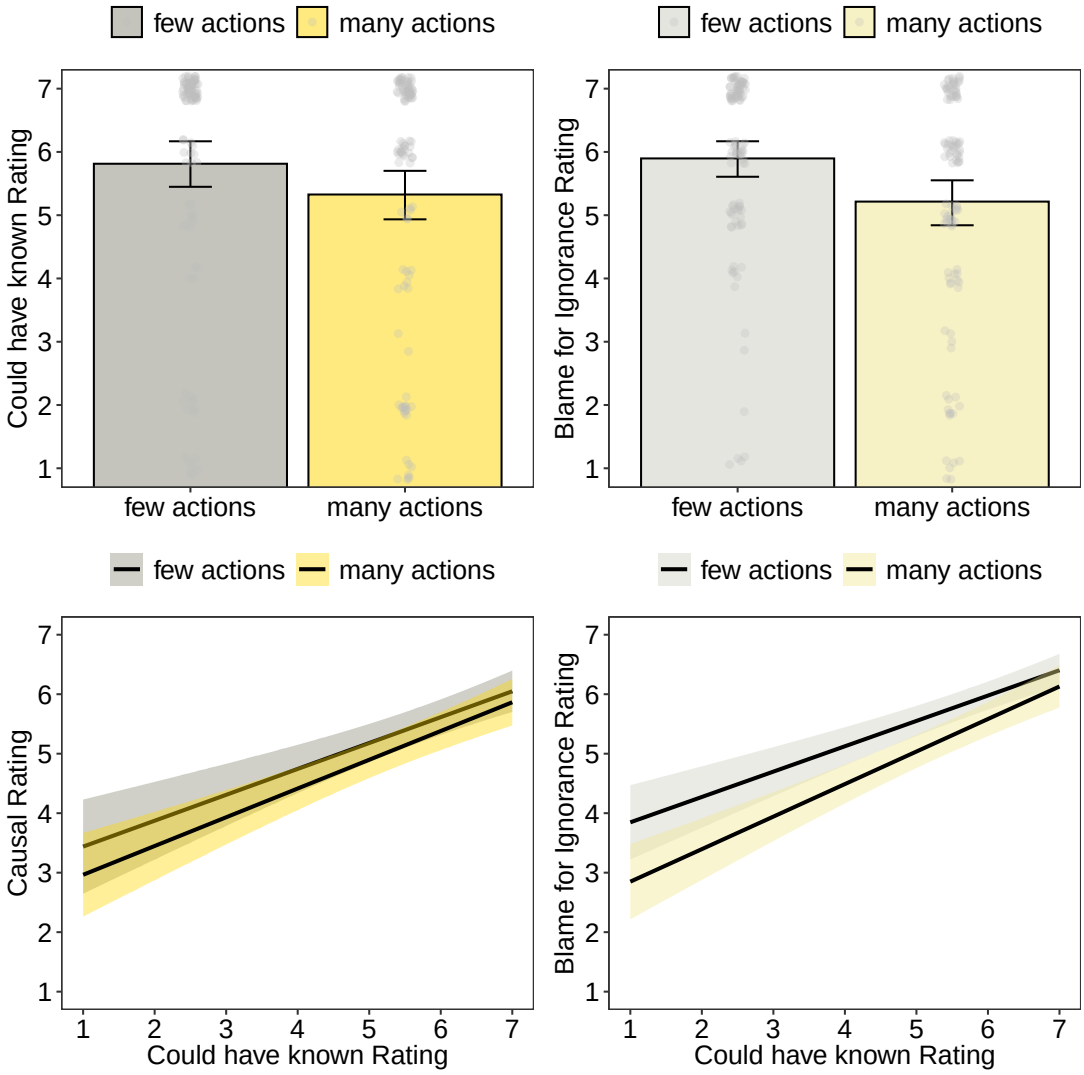

Figure 14. **Experiment 3: Knowledge and Blame Ratings**. Bar graphs depict means of i) "Could have known" ratings and ii) "Blameworthiness for Ignorance" ratings. Regression plots depict "Could have known" ratings as predictor for i) Causal ratings and ii) Blame for Ignorance ratings

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

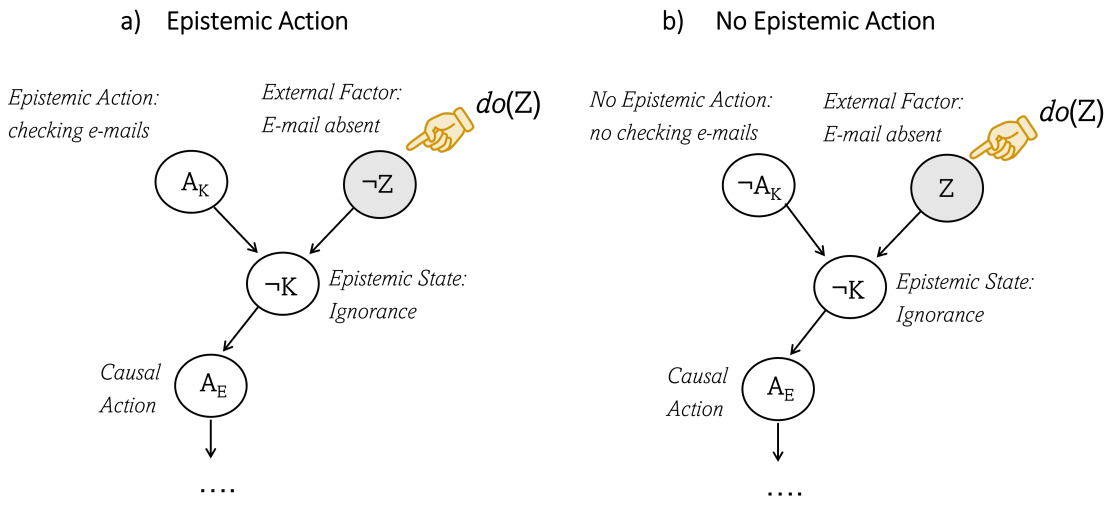

Figure 15. **Experimental Conditions of Experiment 4**. In the 'Epistemic Action' condition (a), the agent performs an epistemic action (reading e-mail), but the external conditions for information acquisition are not given (information not included in the e-mail) and they remain ignorant. In the 'No Epistemic Action' condition (b), 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.

requirements of epistemic actions and epistemic effort might help to show these differences as between-contrasts as well.

The final experiment in this paper 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.

Experiment 4

In our last experiment, we aimed to test whether people take into account whether 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 15a), or whether they do not even perform the epistemic action (see Figure 15b).

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 $N = 133$ ($M_{\text{age}} = 38.36$, $SD_{\text{age}} = 11.38$, $N_{\text{female}} = 57$, 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.

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 2 and 3, in both conditions the agent unwittingly applies the harmful item with negative consequences.

Knowledge and Blame for Ignorance. These ratings and response measures were obtained as in Experiment 2 and 3.

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.

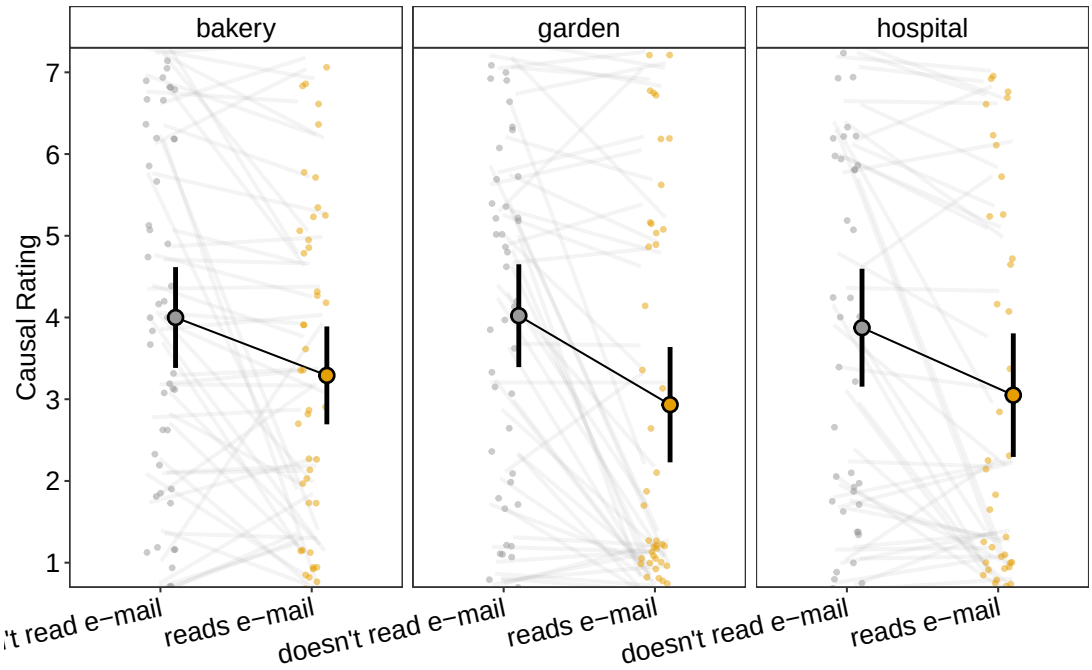

Figure 16. **Experiment 4: Causal Ratings**. Big dots are group means. Error bars depict 95% Confidence Intervals. Small dots connected by lines are individual participants' judgments, jittered for visibility.

Results

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 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 16).

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 or by suggesting an alternative 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..."*) ($N = 90$),
- **"... by reading the e-mail"** ($N = 4$), the agent reading the e-mail (*"If Anne would have read the e-mail"*).

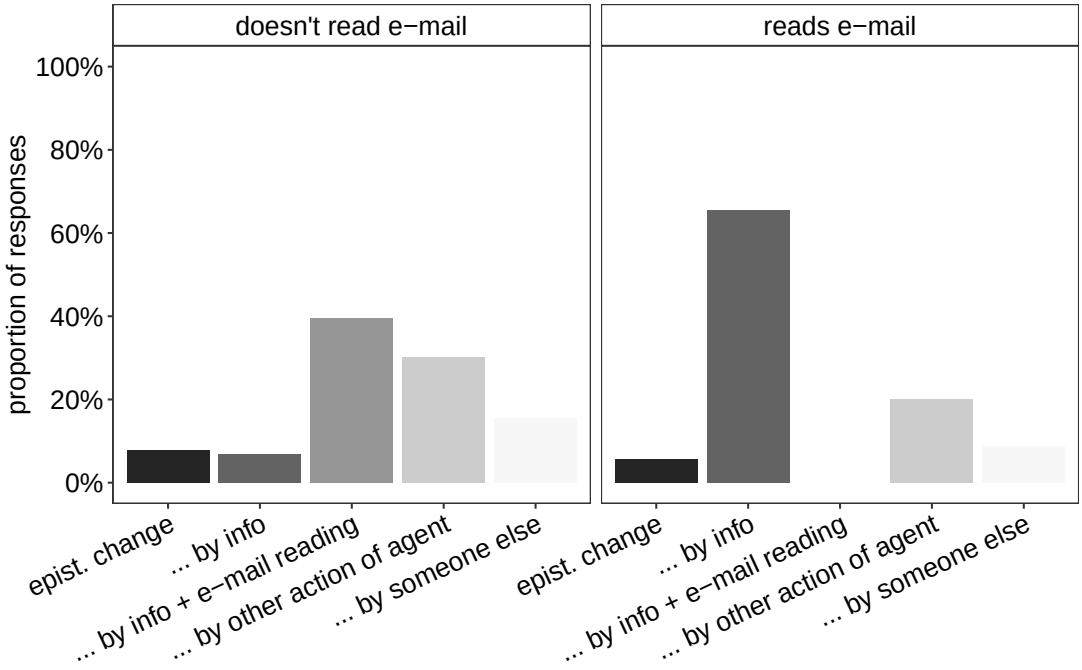

Figure 17. **Experiment 4: Counterfactual responses**. Proportion of responses in the five counterfactual response categories ("epistemic state change, unspecified", "epistemic state change by addition of information in e-mail", "...by addition of information and agent reading e-mail", "...by another epistemic action of the agent" and ".... by another agent")

- **"... by 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.

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*: $\chi^2(4) =$

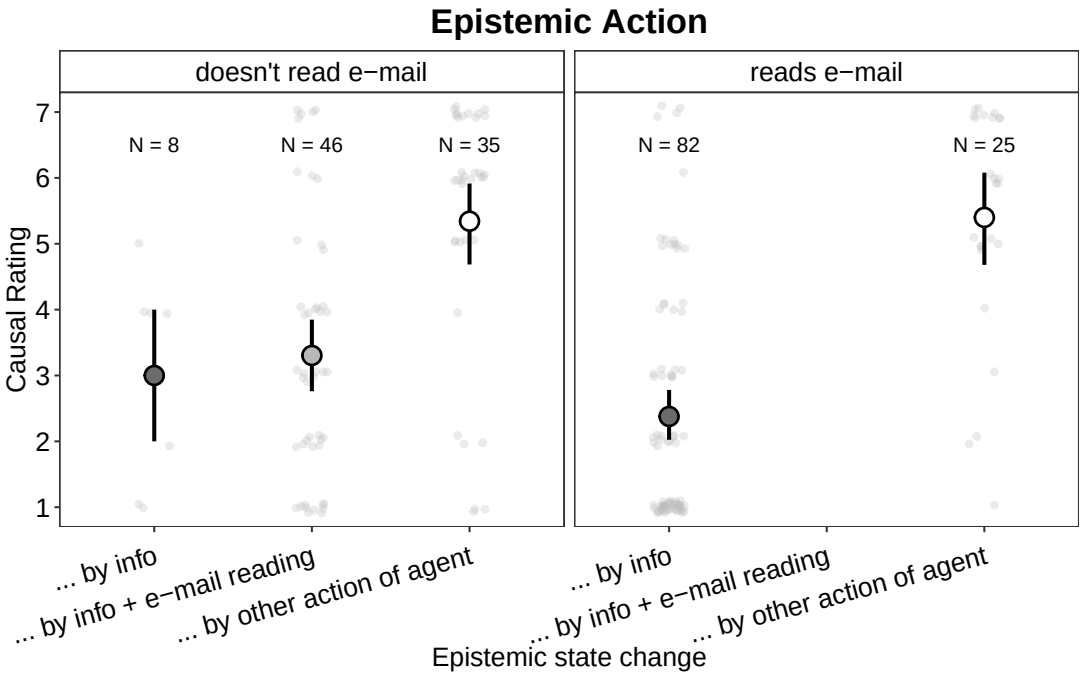

Figure 18. **Experiment 4: Causal Ratings by Counterfactual Response Category**. Causal Ratings of participants who gave imagined the agents' epistemic state change to be caused by i) the e-mail containing the relevant information, ii) the e-mail containing the relevant information and the agent reading it and iii) by some alternative epistemic action by the agent, split by ignorance condition

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 17) (b = -1.80, OR = -.16, SE = .91, z = -1.96, $p < .001$).

Counterfactuals: Sub-group Analysis. As in the studies before, we aimed to analyse 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, $p <$

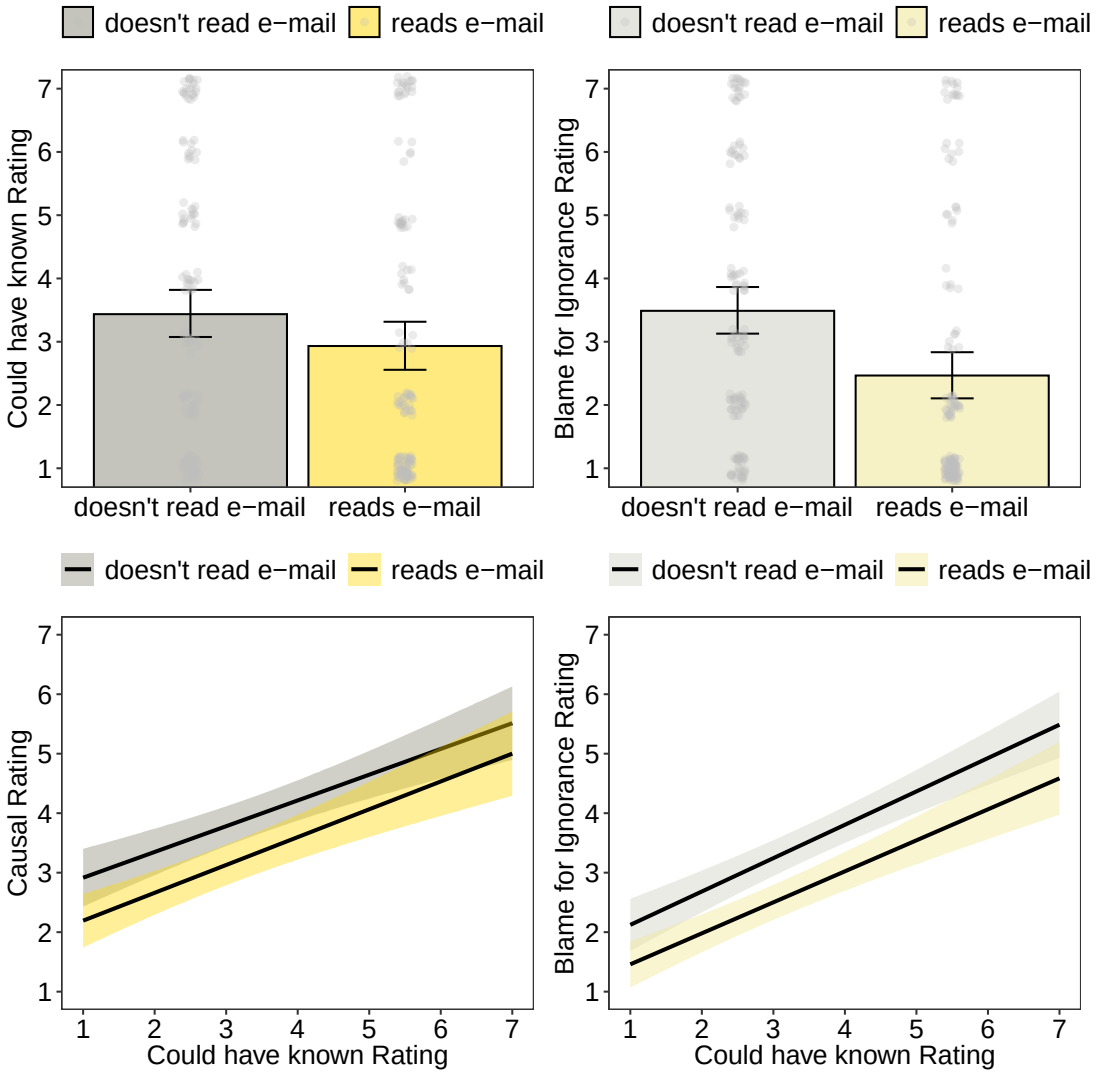

Figure 19. **Experiment 4: Knowledge and Blame Ratings**. Bar graphs depict means of i) "Could have known" ratings and ii) "Blameworthiness for Ignorance" ratings. Regression plots depict "Could have known" ratings as predictor for i) Causal ratings and ii) Blame for Ignorance ratings.

.01] (see Figure 18). 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, $p < .001$].

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 19).

Figure 20. **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).

The agent who did not read the e-mail containing missing information ("not information-seeking") 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])

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*: $F(1) = 72.96$, $p < .001$].

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 20). 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]).

Discussion

In final experiment of this paper, 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 behaviour 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 hypothesis. Based on the agents' prior epistemic (non)-actions, people predicted the agent who had read the e-mail before to do so again. In consequence, people judged the likelihood of a similar future outcome to be lower for the agent who undertook an epistemic action in previous scenarios. The likelihood of the future outcome here includes some natural uncertainty about whether the future epistemic state change also results in the absence of the causal action. It is, however, significantly predicted by the agent's 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 the outcome on epistemic actions under different epistemic contexts.

One might argue that people generally blame the agent in the "did not read e-mail" condition for not reading the e-mail or their negligence, despite the e-mail lacking crucial information. We think the reason why people assign blame as a matter of principle here is precisely because they evaluate the agent's epistemic non-action against a variety of different counterfactual alternatives in which this epistemic non-action would in fact have left the agent in ignorance about pivotal information. This prediction marks a difference between our account and general blame-oriented accounts of causation (Alicke et al., 2012; Sytsma, 2019). Any norm violation that is not of particular relevance for the agent's epistemic states (e.g. forgetting to put the doctor's coat on) should not have an impact on causal or blame judgments about the causal agent for the unknown outcome.

General Discussion

Did Romeo cause Juliet's death? According to recent studies in causal cognition, Romeo's ignorance of the consequence of his actions not only reduces his moral responsi-

bility for Juliet's death, but also his perceived causal contribution. 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 paper, we have investigated the role that agent epistemic states play for causal judgments. We put forward the hypothesis that the influence of ignorance on causal judgements shows how people think about social causation. Specifically, the proposal we make in this paper 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 people's causal judgements about ignorant causal agents map on to 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 changed their epistemic state in the actual world, but would have acquired knowledge in an alternative world under 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 discuss our results with reference to other causal theories and frameworks. 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. Specifically, we want to touch upon some potential implications that we think this work could have for the debate around the relationship between causality and blame, and relatedly, normative vs. non-normative accounts of causation. In the final section, we reflect on the broader implications of the work in this paper.

Integrating Epistemic States into Formal Frameworks of Causation

Halpern & Hitchcock's Means Rea and Possible World Ordering

Halpern and Hitchcock (2015) incorporate defaults, typicality, and normality into their formal causal framework by 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*. Halpern and Hitchcock (2015) propose that possible worlds can be 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. Halpern and Hitchcock (2015) also sometimes refer to "Bob's malice" or "Anne's negligence" as the cause of the fire. In general, 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 (epistemic) 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 properties, i.e. normality and optimal target of intervention, might often align). Crucially, 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.

Structural Model Account of Blame

Chockler and Halpern (2004) extend the counterfactual framework of causation for a definition of blame that takes into account an agent's epistemic state. According to this account, blame is relative to an agent's epistemic state, which is taken to be a set of scenario that the agent considers possible before the action is performed, together with the likelihood of each scenario. An agent's blameworthiness for an action is hence their expected degree of responsibility for the outcome summed over all possible scenarios. 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. In contrast, 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 amounts to $\frac{1}{3} \times \frac{1}{2}$ (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, and is independent of which causal condition A, B, or C was actually present. Chockler and Halpern (2004) argue that people consequently update their blame judgment about the doctor after the

treatment, when her knowledge about whether and under which circumstances the outcome occurred changes. 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 causal 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 of integrating degrees of epistemic uncertainty rather than just binary epistemic states like knowledge vs. ignorance.

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 directly proportional to the normality of a counterfactual (Icard et al., 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Phillips, Morris, & Cushman, 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) that are sensitive to the causal agent's epistemic states. 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.

Process-theory of Causality: Force Theory of Causation

The force dynamics model characterises 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). 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 a physical force. An agents' intention towards a certain state or goal can align or conflict with (mental) force configurations, characterising the exact causal causal relationship between them (E.g. "The police officer enabled vs. prevented the woman from going to the other side of the road."). In contrast to intentions or desires, epistemic states like

knowledge 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.

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 paper posits 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, 2012; Alicke et al., 2012; Danks, Rose, & Machery, in press; Samland & Waldmann, 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., 2014; 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, 2012) 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, 2016a), and will hence be influenced by factors that are relevant for the 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 considerations. Given the influence of our epistemic manipulations on judgements of blameworthiness for ignorance, such a concern is warranted.

The Link Between Counterfactuals and Blame

The account we aim to give here applies to the causation of outcomes of different valences, and we have argued earlier in this paper why epistemic states — and counterfactual interventions on epistemic states — should influence 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 (Alicke & Rose, 2012; Sytsma, 2020b), 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 when and how these normative judgments factor into the counterfactual reasoning process.

Counterfactuals before Blame. One line of response is that the kind of counterfactuals people consider determine their attributions of blame (Sher, 2009), and in consequence causal judgments. 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 systematically refer to counterfactuals containing the (im)mutability of epistemic states ("I couldn't have known that ..."). On such a view, people's causal judgements are an indirect result of how the counterfactuals they imagine influence their attributions of blame and responsibility. Some proponents of a normative account of causation have in fact been open to the idea that counterfactual reasoning influences the moral judgements they take to blend with judgements about causation (see e.g. Sytsma, 2020b).

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 points of norm-deviance. The idea that moral evaluations influence counterfactual thinking, which in turn affects judgments about non-moral properties, is not new (Knobe, n.d.; Phillips, Luguri, & Knobe, 2015; Phillips et al., 2019). Studies demonstrate that moral judgments about normative 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 propose 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; we argue, however, that people might intervene on epistemic states independent of how 'abnormal' the agent's state of ignorance is. As a result, we hypothesise that interventions on ignorance drive people's causal judgements, even if agent's epistemic state of ignorance is not blameworthy or morally bad. Previous studies show that ignorance moderates the effect of normality on causal judgments (Kirfel & Phillips, 2021). The results from our experiments demonstrate that ignorance about the general consequences of one's action has an impact, too.

Pragmatic Effects: Conditioning on Actions

In a series of compelling studies, Samland and Waldmann (2016a) 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 by 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?"). Samland and Waldmann (2016a) find that people's responses on these measures are *not* affected by the 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, Samland and Waldmann (2016b) take their results as evidence for the pragmatic influences on standard causal test questions. 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?") might not be affected by the agent's epistemic states, since they direct people's counterfactual intervention specifically on the causal action. Our account applies to "causal agents", broadly construed.

Bridging Causality and Blame

In sum, we do not aim to rule out that 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 counterfactual reasoning blame judgments come in. The 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 unchanged.

Ultimately, we think that the function of epistemic states in causal cognition that we have been arguing for in this paper might have the potential to shed new light on the debate around the relationship between causal and normative judgments. What causal agents know, intend or believe forms an important starting point for various moral attributions about their behaviour: blame (Alicke, 2008), responsibility (Sytsma, 2019), or accountability (Samland et al., 2016). 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, and further contextual information such as mental states or normative features inform our evaluative response. If anything, our work suggests that theories of blame for ignorance might want to lay out their predictions with causal models that include epistemic states (Sloman et al., 2012). As discussed earlier, it has, however, been argued that causality also has an important "forward-looking" function: causal selection pinpoints useful places for future intervention and action (Hitchcock, 2012; Liquin & Lombrozo, 2020; Lombrozo & Carey, 2006). The factors that are singled out in causal judgments should also lead to useful downstream effects (Danks, 2013; Woodward, 2014). Mental states are the vehicle of agent behaviour and their potential consequences, and therefore present suit-

able targets for causal intervention. Even if interventions on mental states might represent more 'fuzzy' interventions than interventions on specific 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.

Reasonable Person and 'Virtuous' Reasoners

Richer agent models and the integration of epistemic states into formal causal models holds the potential to provide formal tools to model different types of 'mens rea' and legal concepts such as the 'reasonable person' standard (Miller & Perry, 2012; Tobia, 2018). Mapping out the epistemic landscape of what a reasonable person knows (and does) with causal models can provide a more fine-grained reasonable person test and identify precise points of deviation. Such reasonable person models can furthermore be informed empirically and updated with data on various epistemic and behavioural priors, allowing for a more quantitative measure of violation of the reasonable person standard (Alicke & Weigel, 2021; Jaeger, 2020).

The approach we suggest here can also be extended to people's reasoning and the integration of evidence when making decisions under uncertainty (Srivastava, 2011). Causal and blame attributions hence might not only be sensitive to ignorance, but also to epistemic states of uncertainty or the process of (evidential) reasoning that has led to an agent's beliefs and actions. Recent approaches to epistemology have stressed the role of epistemic or intellectual virtues in the evaluation of knowledge states (Greco, 1993; Sosa, 2007). We believe that the account sketched in this paper might be a promising starting point to connect recent theoretical developments in virtue epistemology with the formal frameworks of causal modelling and empirical research on cause and blame judgments (Hundertmark & Kindley, 2021).

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 judge the causal strength of causal factors. While there are many things in the story of Romeo and Juliet that could have taken a different course, even in the final scenario, it seems that their death can 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. In consequence, Juliet would not have stabbed herself out of grief for his death. 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 paper, 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 a causal scenario. Taking into account mental states might explain why we attribute lesser causality to ignorant agents, but also acknowledge 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 paper to provide a first answer to the question *why* this is the case.

Appendix A, Scenarios

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."

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`.

Appendix B, Between Subject Results

Experiment 1, Between Subject Factor Analysis

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]), $t(115) = 1.24$; $p = .22$.

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.

Experiment 2, Between Subject Factor Analysis

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, $F(1) = 0.81$; $p = .44$.

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%).

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, $p = .86$].

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, $p = .01$.

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, $p = .03$.

Experiment 3, Between Subject Factor Analysis

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 ($p = .87$).

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%).

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, $F(1) = 0.51$; $p = .48$.

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, $F(1) = 53.37$; $p < .001$ ($b = 0.28$, $SE = .09$, $t = 3.20$).

Experiment 4, Between Subject Factor Analysis

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]).

Counterfactual Responses. The information acquisition condition significantly predicted people's counterfactual responses, $\chi^2(-4) = 65.49$; $p < .001$, $R^2 = .20$. 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 ($b = -1.80$, $OR = -.16$, $SE = .91$, $z = -1.96$, $p < .001$) (6% vs. 60%).

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, $t(87) = -4.84$, $p < .001$.

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

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), 556–574.

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

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

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

Alicke, M. D., & Weigel, S. H. (2021). The reasonable person standard: Psychological and legal perspectives. *Annual Review of Law and Social Science*, *17*.

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).

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.

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.

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.

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.

Campbell, J. (2007). An interventionist approach to causation in psychology. *Causal learning: Psychology, philosophy and computation*, 58–66.

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues*, *20*, 64–79.

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

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

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

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.

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.

Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of science*, *74*(5), 981–995.

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).

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*, 100785.

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.

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. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Cogsci*.

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

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.

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

Greco, J. (1993). Virtues and vices of virtue epistemology. *Canadian Journal of Philosophy*, *23*(3), 413–432.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). 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.

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. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, *66*, 413–457.

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.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

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., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

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

Hundertmark, F., & Kindley, S. (2021). Making a difference in virtue epistemology. *Synthese*, 1–17.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. Retrieved from `https://doi.org/10.1016%2Fj.cognition.2017.01.010` doi: 10.1016/j.cognition.2017.01.010

Jaeger, C. B. (2020). The empirical reasonable person. *Ala. L. Rev.*, *72*, 887.

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

Kaiserman, A. (2020). Interventionism and mental surgery. *Erkenntnis*, *85*(4), 919–935.

Kasimatis, M., & Wells, G. L. (1995). Individual differences in counterfactual thinking. *What might have been: The social psychology of counterfactual thinking*, 81–101.

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.

Kim, J. (1974). Causes and counterfactuals. *The Journal of Philosophy*, *70*(17), 570–572.

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.

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. (n.d.). *Morality and possibility.* The Oxford Handbook of Moral Psychology. Oxford: Oxford University Press.

Knobe, J. (2009). Folk judgments of causation. *Studies In History and Philosophy of Science Part A*, *40*(2), 238–242.

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

Kominsky, J. F., & Phillips, J. (2019, Oct). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, *43*(11). Retrieved from `http://dx.doi.org/10.1111/cogs.12792` doi: 10.1111/cogs.12792

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kubyshkina, E., & Petrolo, M. (2019). A logic for factive ignorance. *Synthese*, 1–12.

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

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

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.

Le Morvan, P. (2013). Why the standard view of ignorance prevails. *Philosophia*, *41*(1), 239–256.

Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology*, *2*(1), 19–32.

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

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

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Clarendon Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology*, *33*(2), 101–121.

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.

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.

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.

McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of personality and social psychology*, *79*(5), 677.

Meyer, J.-J. C., & Hoek, W. v. d. (1995). Knowledge and ignorance. In *Epistemic logic for ai and computer science* (p. 113–158). Cambridge University Press. doi: 10.1017/CBO9780511569852.005

Miller, A. D., & Perry, R. (2012). The reasonable person. *NYUL Rev.*, *87*, 323.

Miller, S. (2018). Joint epistemic action: some applications. *Journal of Applied Philosophy*, *35*(2), 300–318.

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.

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.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peels, R. (2010). What is ignorance? *Philosophia*, *38*(1), 57–67.

Peels, R. (2012). The new view on ignorance undefeated. *Philosophia*, *40*(4), 741–750.

Phillips, J., Luguri, J., & 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.

Piaget, J. (1965). *The moral judgment of the child.(translated by marjorie gabain)*. Routledge & K. Paul (1965, 1932).

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

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.

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.

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016). 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., & 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. (2016a). 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

Samland, J., & Waldmann, M. R. (2016b). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176.

Sayre, F. B. (1932). Mens rea. *Harvard Law Review*, *45*(6), 974–1026.

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, *114*(3), 327–358.

Shakespeare, W. (1858). *Romeo en julia*. AC Kruseman.

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.

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

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind & Language*, *27*(2), 154–180.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, *66*, 223–247.

Smith, H. (1983). Culpable ignorance. *The Philosophical Review*, *92*(4), 543–571.

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge, volume i* (Vol. 1). OUP Oxford.

Spellman, B. A., & Gilbert, E. A. (2014). Blame, cause, and counterfactuals: The inextricable link. *Psychological Inquiry*, *25*(2), 245–250.

Srivastava, R. P. (2011). An introduction to evidential reasoning for decision making under uncertainty: Bayesian and belief function perspectives. *International Journal of Accounting Information Systems*, *12*(2), 126–135.

Sytsma, J. (2019). The character of causation: Investigating the impact of character, knowledge, and desire on causal attributions. *pre-print*.

Sytsma, J. (2020a). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*.

Sytsma, J. (2020b). Resituating the influence of relevant alternatives on attributions.

Sytsma, J., & Livengood, J. (2019). Causal attributions and the trolley problem. *pre-print*.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive science*, *12*(1), 49–100.

Tobia, K. P. (2018). How people judge what is reasonable. *Ala. L. Rev.*, *70*, 293.

Van Der Hoek, W., & Lomuscio, A. (2004). A logic for ignorance. *Electronic Notes in Theoretical Computer Science*, *85*(2), 117–133.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

Walsh, C. R., & Byrne, R. M. (2007). How people think "if only..." about reasons for actions. *Thinking & Reasoning*, *13*(4), 461–483.

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.

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.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology: General*, *136*(1), 82.

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. (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/`

Yee, T. W., et al. (2010). The vgam package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34.

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. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, *107*(3), 410–426.