Inference from explanation

Lara Kirfel
University College London

Thomas Icard
Stanford University

Tobias Gerstenberg*
Stanford University

Abstract

What do we communicate with causal explanations? Upon being told, "E because C", one might learn that C and E both occurred, and perhaps that there is a causal relationship between C and E. In fact, causal explanations systematically disclose much more than this basic information. Here, we offer a communication-theoretic account of explanation that makes specific predictions about the kinds of inferences people draw from others' explanations. We test these predictions in a case study involving the role of norms and causal structure. In Experiment 1, we demonstrate that people infer the normality of a cause from an explanation when they know the underlying causal structure. In Experiment 2, we show that people infer the causal structure from an explanation if they know the normality of the cited cause. We find these patterns both for scenarios that manipulate the statistical and prescriptive normality of events. Finally, we consider how the communicative function of explanations, as highlighted in this series of experiments, may help to elucidate the distinctive roles that normality and causal structure play in causal judgment, paving the way toward a more comprehensive account of causal explanation.

Keywords: inference; explanation; causal judgment; normality.

*Corresponding author: Tobias Gerstenberg (gerstenberg@stanford.edu), Department of Psychology, 450 Jane Stanford Way, Building 420, Office 302, Stanford, CA 94305

The ability to *explain* is at the core of how we understand the world and ourselves (Craik, 1943; Salmon, 1984; Woodward, 2003). As scientists, we are often not content with merely being able to predict what will happen. Instead, we strive for a deeper understanding of the underlying causal laws or mechanisms that dictate *how* and *why* the world works the way it does. Explanations also play a critical role in our everyday lives (Hagmayer & Osman, 2012; Heider, 1958; Lombrozo, 2006). One of the most important functions of causal explanation is in interpersonal interaction: we understand one another as guided by reasons, with much of behavior intelligible in decidedly causal terms (Buss, 1978; Davidson, 1963; Malle, 1999).

Despite the prominence of explanation throughout human affairs, we still lack a detailed understanding of how exactly explanations work. As many have emphasized (Friedman, 1974; Keil, 2006; Lombrozo, 2006; van Fraassen, 1980), explanations – conceived as answers to "why?" questions – facilitate *understanding* on the part of the individual receiving the explanation. This observation highlights a significant *communicative* dimension of explanation. While theorists have often attempted to study the subject in relative abstraction from concrete discursive contexts (e.g., Lewis, 1986; Salmon, 1984; Strevens, 2008), a number of researchers have argued that many idiosyncratic features of explanation demand that we take this communicative dimension more seriously (Achinstein, 1983; Hilton, 1990; Potochnik, 2016; Turnbull & Slugoski, 1988). In this light the main question becomes: How exactly do we employ causal explanations to impart understanding? That is, what kinds of strategies do we use to produce and interpret causal explanations?

This question is especially acute since, on the face of it, we seem to say very little explicitly when offering explanations (cf. Hemmatian & Sloman, 2018, for a striking example, where mere labels are used as explanations). A prototypical causal explanation may involve nothing more than a specification that "E happened because C happened," essentially just citing two events, C and E. As Keil (2006) frames the issue, "Somehow, people manage to get by with highly incomplete or partial explanations of how the world around them works [. . .] We have yet to understand the nature of such compressions of information" (p. 135). Answering this question promises to illuminate not only a central aspect of human cognition, but also potential ways we may be able to simulate explanatory behavior in artificial agents, a notoriously challenging task (cf. Byrne, 2019; Marcus & Davis, 2019).

Our aim in this article is to establish groundwork for a more detailed account of how people communicate using explanations, and in particular of the subtle but systematic patterns of inference people draw upon receiving an explanation. We focus here on *causal* explanations. The idea that listeners in a dialogue go far beyond what is explicitly (or "literally") said by a speaker is widely appreciated, and there are well-developed theoretical frameworks for studying this capacity (e.g., Clark, 1996; Goodman & Frank, 2016; Levinson, 2000). Situated within this wider theoretical context, we offer an account of explanatory dialogue in particular. How do causal explanations function, such that general communicative principles allow people to learn so much from so little?

As a way into this question, we draw upon a growing body of research on how event normality and causal structure affect people's causal judgments (e.g., Gerstenberg & Icard, 2020; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). Previous experimental work has emphasized systematic patterns in participants' judgments about various causal claims. In the present work we turn this around,

probing not just what explanations people deem reasonable, but whether they can leverage these very intuitions to make appropriate inferences from claims made by others. Specifically, we show in two studies that people are able to infer specific information about *event normality* when provided information about *causal structure*, and conversely, they can infer causal structure when provided information about event normality. These case studies, we argue, underscore much of what is characteristic of explanatory dialogue. People's inferences stem from a combination of generic conversational principles and commonly shared causal intuitions.

After presenting our theoretical proposal in greater depth and deriving several key predictions, we then test these predictions in two main studies. At the end, we discuss our findings in the context of existing accounts of causal explanation and causal reasoning. In particular we suggest that the communicative dimension of explanation highlighted in these experiments may even help to *explain* some of the most prominent patterns in our causal judgments.

Learning from explanations

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

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

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

about important aspects of the situation. Learning from explanations crucially involves communication.

Explanations in communication

Generically, communication involves (at least) two interlocutors with partially overlapping knowledge about the world, and about the meanings of various possible "signals" (Clark, 1996). These two sources of knowledge in concert can facilitate strikingly efficient transfer of information (e.g., Gibson et al., 2019). Part of what makes such efficient information transfer possible is the fact that people generally adhere to systematic discourse patterns, in how they produce and interpret linguistic utterances (Grice, 1975). This allows the meanings of signals to be relatively underspecified, since interlocutors can rely on a combination of world knowledge and tacit understanding of conversational principles to go far beyond what is said literally (Goodman & Frank, 2016; Grice, 1975; Levinson, 2000). We believe that such systematic pragmatic principles are key to the proper analysis of what people manage to learn from causal explanations.

Our proposal thus combines two ingredients: some simple but general principles of communication, and a minimal analysis of what the signal "E because C" means. For a first pass at the latter, we take the meaning to be captured by the circumstances in which it would be appropriate to utter the phrase.[1] These two ingredients then allow us to predict what people will infer from a causal explanation. If a speaker S utters to a listener L, "E because C," then L may think about how the world must have been in order for this to have been an appropriate thing for S to say. Assuming that S is a cooperative speaker, using the phrase in the normal way, and knowledgeable about the relevant state of the world, L will be able to infer that the world must have been that way.[2]

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

[1] Though we do not assume in general that linguistic meaning reduces to circumstances of use or acceptance, there is a prominent tradition of thought arguing for precisely this reduction (e.g., see Horwich, 1998). Such a gloss will be adequate for our purposes here.

[2] Here L would be a "level-1" listener in the terminology of Goodman and Frank (2016), in that L assumes S is simply employing the ordinary meaning of the phrase, and in particular S need not be considering how L might interpret the phrase. Certainly more complex scenarios are imaginable, but for the purposes of this paper such level-1 reasoning will be sufficient.

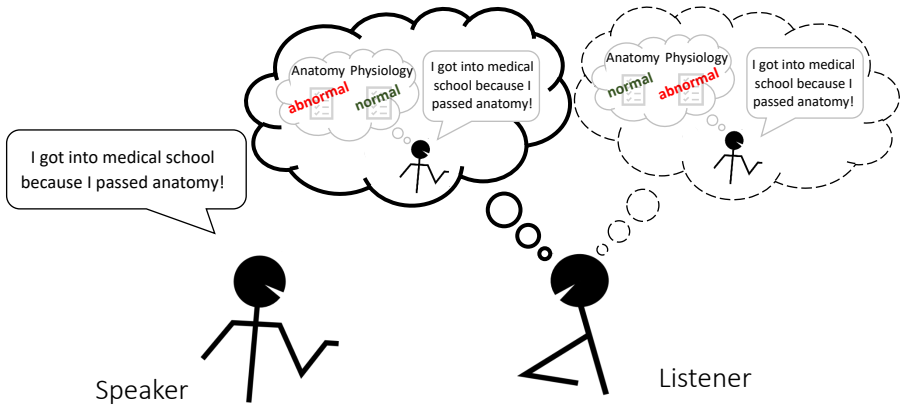

Figure 1

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

to say.[3]

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

The influence of norms and causal structure on causal explanations

When multiple causes contributed to an outcome, people tend to select only a few causes in their explanation of what happened rather than citing all of them. Causal selection

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

moreover follows systematic patterns. For example, people often prefer abnormal over normal events as causes (Cheng & Novick, 1991; Halpern & Hitchcock, 2015; Hart & Honoré, 1959/1985; Hesslow, 1988; Hilton & Slugoski, 1986; Phillips & Cushman, 2017). When two causes are each necessary for producing a certain outcome (conjunctive structure), people judge the abnormal event as more causal (Gerstenberg & Icard, 2020; Icard et al., 2017; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Kominsky et al., 2015). The influence of normality on causal selections has been shown both for *statistical norms* (i.e. the frequency with which an event occurred in the past), as well as *prescriptive norms* (i.e. whether an event adheres to or violates a social or moral norm).

While there is an ongoing discussion on how best to explain this preference for abnormal causes (Alicke, 2000; Kominsky & Phillips, 2019; Samland & Waldmann, 2016; Sytsma, Livengood, & Rose, 2012), recent research has found that when two causes are each sufficient for the outcome (disjunctive structure), people show a preference for the *normal* over the abnormal cause (Gerstenberg & Icard, 2020; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Icard et al., 2017). For instance, Gerstenberg and Icard (2020) had participants watch video clips of physically realistic interactions between inanimate objects (see Figure 2). In these clips, ball A and ball B enter the scene from the right, and are headed toward a stationary ball E. In order to hit ball E, each of them needs to pass through a blocker. Crucially, the blockers differ in how likely they are to let a ball go through. While the light red blocker has a 80% chance of letting a ball go through, the dark red blocker only has a 20% chance. The clips came in two different setups that were manipulated between participants: In the conjunctive setup, both ball A and ball B need to go through the blocker in order to make ball E through the gate. In the disjunctive setup, being hit by either ball A or ball B is sufficient to make ball E go through the gate. Participants watched ten of these clips and learned how likely it was for each blocker to let a ball go through. In the test phase, participants watched a clip in which both balls went through the blocker and, as a result, ball E went through the gate (Figure 2 middle). Participants were asked to select which explanation better described what happened: "Ball E went through the gate because ball A [ball B] went through the motion block".

The results showed – consistent with prior research – that when two causal events were both necessary to make the outcome happen, participants selected the abnormal event (i.e. the ball that was unlikely to go through the blocker). However, when either of two events was individually sufficient, participants selected the normal event (i.e. the ball that was likely to go through the blocker). This effect is surprising because unlike what has been assumed for decades (e.g. Hart & Honoré, 1959/1985), people don't show a uniform preference for abnormal causes. Instead, event normality and causal structure interact to determine causal selections.

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

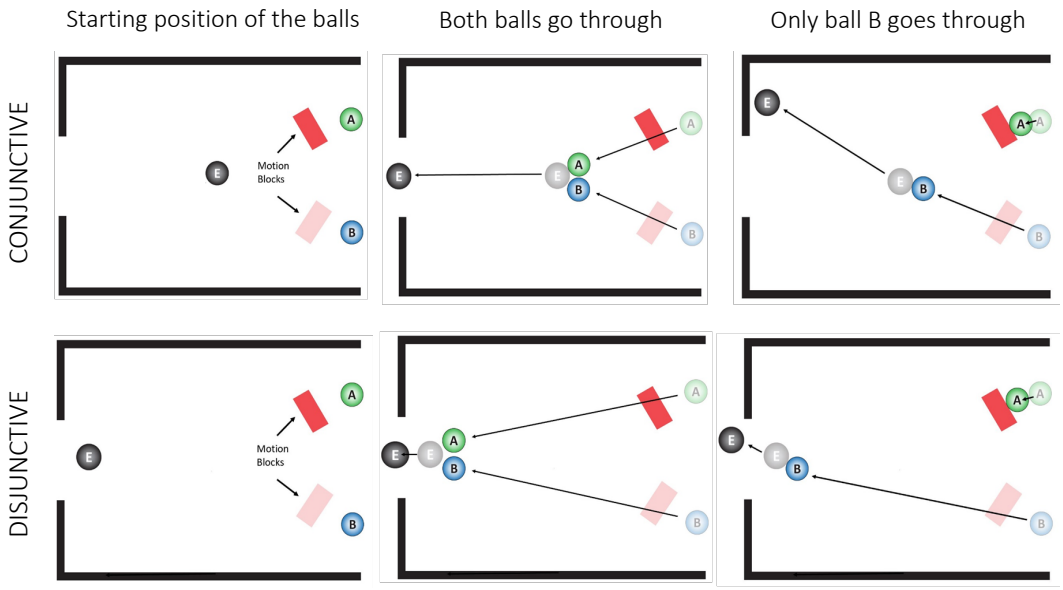

Figure 2

Diagrammatic illustration of clips used in Gerstenberg and Icard (2020) (original clips varied slightly). The top row shows the conjunctive causal structure, the bottom row the disjunctive structure. The color of each blocker indicates its probability of blocking a ball. The dark red blocker has an 80% chance of blocking a ball, and the light red blocker has a 20% chance of blocking a ball. The first column shows the starting position of the balls, the second a case in which both balls went through the blocker, and the third column a case in which only ball B went through the blocker.

argues that norms affect causal judgments by influencing how blameworthy or responsible an action was (Alicke, 2000; Alicke & Rose, 2012; Samland & Waldmann, 2016; Sytsma & Livengood, 2019). People select causes that are more deserving of blame. Finally, there is work in both philosophy and psychology arguing that causal selections point out suitable targets for intervention (Hitchcock, 2012; Lombrozo, 2010; Woodward, 2003). People select causes that are likely to make a difference to the outcome in the future (Gerstenberg & Icard, 2020; Hitchcock & Knobe, 2009).

Our concern in the present work is not primarily to offer a new account of how or why people's causal selections show these particular patterns, nor do our experiments aim to adjudicate among the accounts sketched above. Rather than just demonstrating that people take into account norms and causal structure when making causal judgments or giving explanations, we aim to test whether people also reason about these considerations in analogous ways when drawing inferences from explanations. More precisely, we capitalize on these systematic patterns of causal judgment to help address our main question: how do people manage to communicate so much with such partial and compressed explanations? Other factors that influence causal judgments, such as the temporal order of events, could similarly serve this purpose (see Henne, Kulesza, Perez, & Houcek, 2021). In the General

Discussion, we will discuss how our framework and results relate to the different theories of what drives people's causal selection preferences.

Predictions

We propose that the inferences people draw from others' explanations can be predicted on the basis of general principles of conversation together with an accurate construal of what people take claims of the form "E because C" to mean. We assume the large body of research on causal judgment – including on the roles of norms and causal structure – offers a suitable hypothesis about the latter. This broad proposal issues in a number of concrete predictions, which we outline below.

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

From prior research, we know that both the normality of causes as well as the underlying causal structure influence causal judgments and explanations (Gerstenberg & Icard, 2020; Icard et al., 2017; Kominsky et al., 2015). As a first hypothesis, we predict a replication of these effects in our experiments. Specifically, we predict that when an abnormal and a normal cause bring about an outcome E, people will tend to select the abnormal cause as an explanation for why E happened when both causes were needed ("conjunctive causal structure"), but will tend to select the normal cause when either cause would have been sufficient ("disjunctive causal structure"; Figure 3a). While most prior work has found these kinds of effects on continuous causal judgments (e.g. Icard et al., 2017; Kominsky et al., 2015), we predict that the same pattern will hold for discrete causal selections (see also Gerstenberg & Icard, 2020). Furthermore, we wanted to replicate these effects because most prior work has used written vignettes whereas in our experiments we show participants animated causal scenarios.

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

When a causal explanation is given and the causal structure is known, we predict that people can infer the corresponding normality of the cited cause $p(\text{normality of the cited cause} \mid \text{causal structure})$ (see Figure 3b). More precisely, knowing that there are only two possible options, they can infer whether the cited cause was abnormal or normal. We assume that people make this inference by considering what they themselves would have said in the given situation. For instance, consider the example in Figure 1. Here, the listener knows that the structure is conjunctive – the speaker needed to pass both anatomy and physiology to get into medical school. When the speaker states that she got into medical school because she passed anatomy, the listener infers the event normality by considering how likely she would have said the same thing in the two possible situations. Given the general preference for citing abnormal causes in conjunctive situations, the listener infers that passing anatomy was abnormal. A speaker would be more likely to cite passing anatomy as the cause when this event was abnormal compared to when it was normal. More generally, if the causal structure is conjunctive, participants will infer

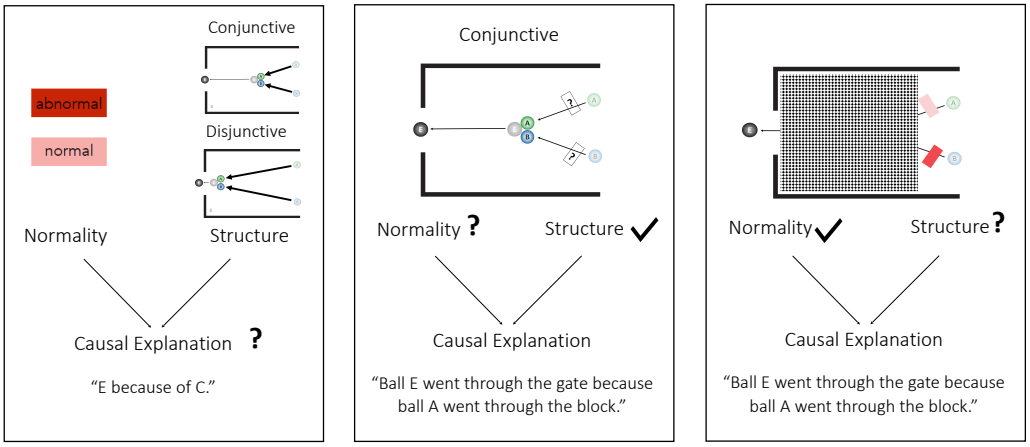

Figure 3

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

that the cited cause is likely to be the abnormal event. In contrast, if the causal structure is disjunctive, participants will infer that the cited cause is likely to be the normal event.

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

We predict that people can infer the causal structure of the situation based on what they know about the normality of the cause cited in the explanation (see Figure 3c). Again, this prediction rests on the assumption that people make this inference considering two concrete hypotheses according to which the causal structure is either conjunctive, or disjunctive. Concretely, we propose that participants solve this communication problem by inferring a conditional probability $p(\text{causal structure} \mid \text{normality of the cited cause})$. We hypothesize that this probability is naturally deconstructed using Bayes' rule:

$$p(\text{structure} \mid \text{normality}) = \frac{p(\text{normality} \mid \text{structure}) \cdot p(\text{structure})}{\sum_{i=1}^{n} p(\text{normality} \mid \text{structure}_i) \cdot p(\text{structure}_i)} \quad (1)$$

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

abnormal event for conjunctive structures, and the normal event for disjunctive structures, we predict that most participants will infer a conjunctive structure if the abnormal event is cited as the cause, and a disjunctive structure if the normal event is cited.

It is worth pausing to consider how striking this hypothesis is. For instance, returning to our example of Suzy, the hypothesis predicts that if a listener knew which test was going to be difficult but did not know whether passing both tests was necessary for admission, or whether just one of the two tests would be sufficient, the listener would be in a position to infer which of these was the case from Suzy's utterance.

Hypothesis 4: Individual differences in inferences from explanations

In the general case, we expect that listeners take into account their knowledge of the speaker when interpreting the speaker's explanations (Goodman & Stuhlmüller, 2013; Kamide, 2012; Schuster & Degen, 2019; Yildirim, Degen, Tanenhaus, & Jaeger, 2016). For example, if a listener happened to know that a speaker has a general tendency to cite abnormal events as causes no matter what the causal structure is, then the listener wouldn't be able to infer the causal structure when the speaker cited an abnormal cause. In the settings that we consider, listeners don't have any speaker-specific information. Accordingly, we assume that listeners will consider what explanation they themselves would have given.

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

In the following, we will report two experiments testing these hypotheses. We test Hypothesis 1 in both experiments. Additionally, Experiment 1 tests Hypothesis 2 and Experiment 2 tests Hypothesis 3. For both experiments, we will look at aggregate results, but also analyze the data by taking into account interindividual differences. Both experiments use two types of norm violations: *statistical norm* violations (using the billiard ball setup shown in Figure 2), and *prescriptive norm* violations involving a scenario with intentionally acting agents (Figure 4).

Experiment 1: Inferring normality given causal structure

In Experiment 1, we test whether participants can infer the normality of an event cited in an explanation based on knowledge about the causal structure of the situation.[4]

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

Methods

Participants and Design

We recruited 210 participants ($\text{Mean}_{\text{age}} = 33$, $\text{SD}_{\text{age}} = 9$, $N_{\text{female}} = 77$, $N_{\text{non-binary}} = 2$, $N_{\text{undisclosed}} = 4$) via Amazon Mechanical Turk (Crump, McDonnell, & Gureckis, 2013). 56 participants were excluded for failing one or more exclusion criteria specified below, leaving a final sample size of $N = 154$ (26.7% excluded). The experiment has a 2 causal structure (conjunctive vs. disjunctive) × 2 norm type (statistical vs. prescriptive) design. Both factors were manipulated between participants. Participants were randomly assigned to the four separate conditions, *statistical normality & conjunctive structure* ($N = 30$), *statistical normality & disjunctive structure* ($N = 37$), *prescriptive normality & conjunctive structure* ($N = 46$), and *prescriptive normality & disjunctive structure* ($N = 41$). All studies reported in this paper were approved by Stanford's Institutional Review Board (IRB-48665).

Statistical Normality: Selection Task

We closely followed the paradigm in Gerstenberg and Icard (2020). Participants were informed that they were going to see video clips of colliding billiard balls, followed by a diagram and description of the billiard ball setup (see Figure 2).

In the *conjunctive* condition, participants saw a diagram illustrating that both balls A and B needed to go through the blockers in order for ball E to go through the gate (see Figure 2, "Conjunctive"). In the *disjunctive* condition, participants, were informed that either ball A or ball B's going through the blockers is sufficient for ball E to go through the gate (see Figure 2, "Disjunctive").

In our experiment, participants were informed that the position of the two blockers may vary from scene to scene. In some setups, the light red blocker would be at the top, while in others, the light red blocker would be at the bottom. Subsequently, participants were asked a series of comprehension check questions about the billiard ball setup.[5]. In order to be included in the experiment, participants had to answer all check questions correctly, and they were given three attempts to read through the instructions and answer the check questions correctly. If a participant failed to answer the check questions after the third attempt, they were forwarded to the end of the study and thanked for their participation.

Participants who answered all of the comprehension check questions correctly then continued with a task in which they themselves selected a causal explanation. This task served two purposes. First, to further familiarize participants with the scenario. Second, to acquire data on participants' own explanation preferences. Participants first only viewed the beginning of the clip. The clip paused shortly after ball A and B entered the scene. Participants were asked to what extent they agreed with the following three predictions: (1) "Ball A will hit ball E.", (2) "Ball B will hit ball E.", and (3) "If only one of the two balls goes through the block and hits ball E then ball E will go through the gate." Participants provided their responses on sliding scales with the end points labeled as "not at all" (0) and "very much" (100). The position of the blockers was counterbalanced across participants.

[5]For example, participants were shown a diagram of a situation and then asked: "In this set-up, if only one of the balls go through the motion block and hit ball E, ball E will go through the gate" with the response options being true/false. See the materials posted online for the full list of comprehension check questions: `https://github.com/cicl-stanford/inference_from_explanation`

We only included participants in the final analysis who rated the chance of the normal billiard ball to hit Ball E higher than that of the abnormal ball, and who responded < 50 in the conjunctive or > 50 in the disjunctive condition for statement (3). These attention check questions made sure that participants had correctly encoded the information in the instructions. The clip then continued to play. Both balls went through the blocker and ball E went through the gate. Participants were asked to select which of the following two statements better described what happened: "Ball E went through the gate because ball A / ball B went through the motion block." We used a two-alternative forced-choice task (rather than a continuous judgment) to match the explanation format that participants received later in the test phase.

Statistical Normality: Inference Task

In the final inference task, participants received a diagram showing a conjunctive (or disjunctive) billiard ball setup in which both ball A and ball B went through the blocker and ball E went through the gate. However, the causal diagram did not include any information about the normality of the two blockers (see Figure 3b). Participants were then told that Ben, a fictional participant, had witnessed the depicted scene and, as in the selection task before, had been asked to select an explanation that best explained what happened. Participants were told that Ben selected the explanation "Ball E went through the gate because ball A [ball B] went through the blocker." We counterbalanced which ball Ben's explanation referred to (ball A or ball B).

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

Prescriptive Normality: Selection Task

To manipulate prescriptive normality, we created an animated video version of the "motion detector" vignette from Kominsky et al. (2015). In this vignette, Suzy and Billy work on a project of national security and they both share an office. This office has a motion detector. In the *conjunctive* condition, the motion detector goes off if more than one person enters the office (Figure 4a, Conjunctive). In the *disjunctive* condition, the motion detector goes off as soon as one person enters the office (Figure 4a, Disjunctive).

Figure 4

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

Given the confidentiality of the project, it is sometimes required that one employee works alone in the office. As a result, on certain days the company's boss instructs both employees that either only Suzy or Billy should come into the office at 9am the next morning, while the other one is supposed to stay away from the office (Figure 4 b). Who is instructed to come in and who to stay away may vary from day to day. Participants were provided with written instructions and the diagrams in Figure 4, followed by four comprehension questions that they needed to answer correctly before proceeding.

In both conditions, participants then saw a video, separated into two parts, showing one morning in the conjunctive [disjunctive] office, and what happened the day before. In

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

Prescriptive Normality: Inference Task

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

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

Results

Figure 5a shows participants' causal selections as a function of the causal structure of the scenario (conjunctive vs. disjunctive) and the type of norm that was manipulated (statistical vs. prescriptive). Table 1 shows the results of a Bayesian logistic regression

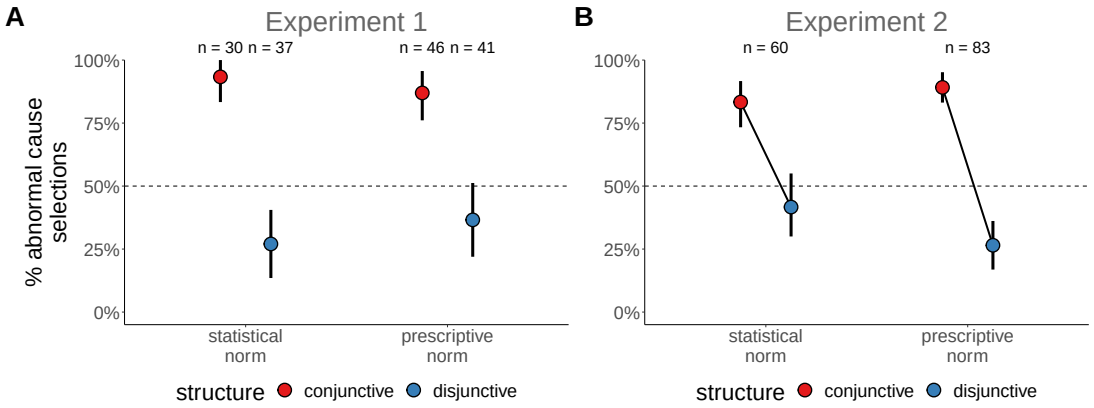

Figure 5

Participant's causal selections in a) Experiment 1 and b) Experiment 2. The data points show the percentage of participants selecting the abnormal cause as a function of causal structure (conjunctive or disjunctive) and norm type (statistical or prescriptive). In Experiment 2, each participant made a choice for both structures as indicated by the lines connecting the data points. The causal selection task replicates and extends the pattern of causal selections from previous studies (Gerstenberg & Icard, 2020). Note: Error bars are bootstrapped 95% confidence intervals.

model of participants' selections.[6] Selections differed as a function of the causal structure. Participants were more likely to select the abnormal cause for conjunctive causal structures (89%) compared to disjunctive structures (32%). There was no effect of type of norm on participants' selections, and no interaction effect between structure and norm.[7]

Table 1

Experiment 1 – Causal selection*: Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian regression model.* Note*: The units are log odds. The dependent variable (selection) was coded as 1 = abnormal cause, 0 = normal cause.*

model specification: `selection ~ 1 + structure * norm`

term	estimate	lower 95% CI	upper 95% CI
intercept	0.81	0.34	1.34
structure	1.60	1.11	2.13
norm	0.10	−0.39	0.64
structure:norm	0.34	−0.20	0.86

[6] All Bayesian models were written in Stan (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019). We also report frequentist statistical analyses in the appendix as well as in the online materials here `https://cicl-stanford.github.io/inference_from_explanation/`.

[7] All categorical predictors were coded using sum contrasts. We adopt the convention of calling something an effect if the 95% highest density interval (HDI) of the estimated parameter in the Bayesian model excludes 0.

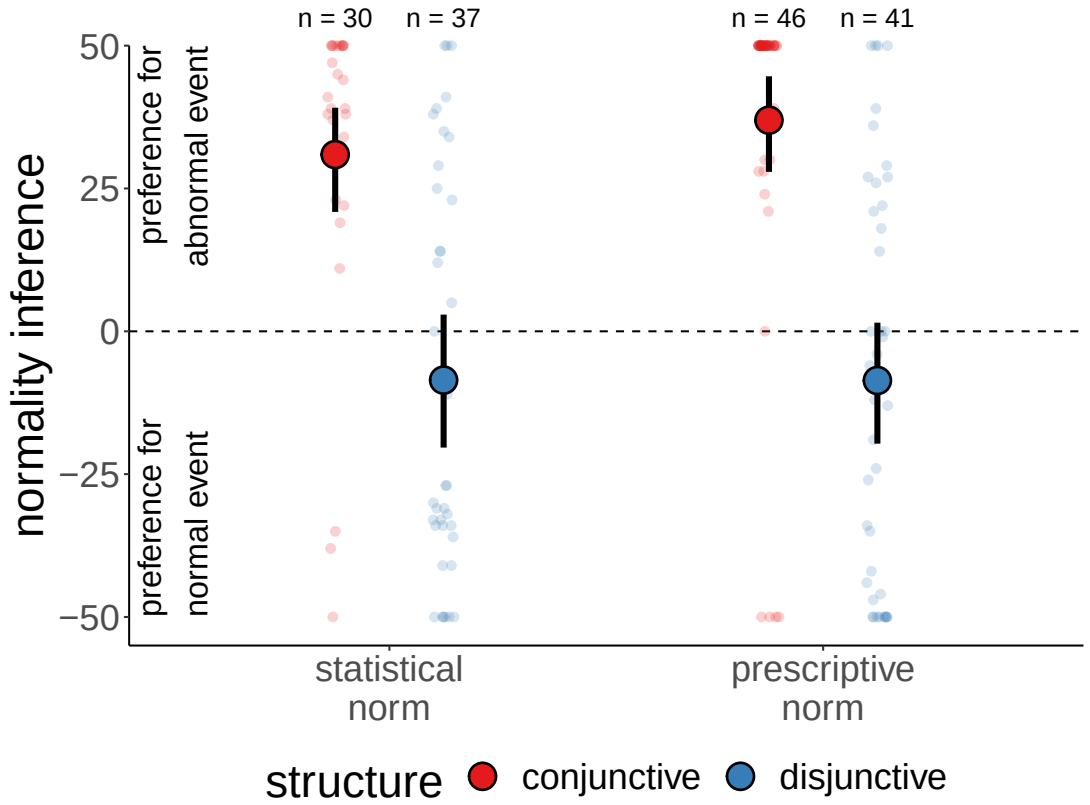

Figure 6

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

Figure 6 shows participants' inferences about the event's normality as a function of causal structure and type of norm. Causal structure strongly affected participants' judgments (see Table 2). Participants inferred that the event cited in the explanation was abnormal when the causal structure was conjunctive (Mean $= 84.58, \text{SD} = 28.13$), and normal when the structure was disjunctive (Mean $= 41.42, \text{SD} = 34.99$). Note that people were more certain that the cited cause was abnormal in the conjunctive structure than that it was normal in the disjunctive structure. The norm manipulation (statistical vs. prescriptive) had no effect on participants' inferences, and there was also no interaction effect between structure and norm type.

As predicted, we found a close correspondence between participants' causal selections and the inferences they drew about the normality of a cause given knowledge about the causal structure. This correspondence becomes even clearer when one compares the proportion of participants who selected the abnormal cause as a function of the causal

structure (as shown in Figure 5) to the proportion of participants who had a preference for the abnormal event in their normality inference. To determine the latter, we simply calculated the proportion of participants who exhibited a preference for the abnormal cause (i.e. whose judgment was greater than 0 in Figure 6). Table 3 shows that there is a very close correspondence between the percentage of participants who selected the abnormal cause as a function of the type of norm and the causal structure (selection), and the percentage of participants who inferred that a selected cause was abnormal (inference).

Individual differences

The tight relationship between causal selections and normality inferences is also demonstrated by breaking down participants' normality inferences based on whether they themselves selected the abnormal or normal cause as a function of the causal structure (see Figure 7). Generally, participants tended to interpret an explanation in line with what they themselves would have said in the same situation. Interestingly, participants who themselves selected the abnormal cause in the conjunctive condition, had a stronger preference to infer the abnormal event than participants who selected the abnormal cause in the disjunctive condition. Moreover, as already apparent from Figure 6, there is an asymmetry in participants' inferences. Participants who selected the abnormal cause in the

Table 2

Experiment 1 – Normality inference*: Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian regression model. Note: For the dependent variable (normality rating), 100 = abnormal and 0 = normal.*

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

term	estimate	lower 95% CI	upper 95% CI
intercept	62.83	57.57	68.11
structure	21.22	16.27	26.35
norm	−1.47	−6.37	3.83
structure:norm	−1.46	−6.41	3.53

Table 3

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

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

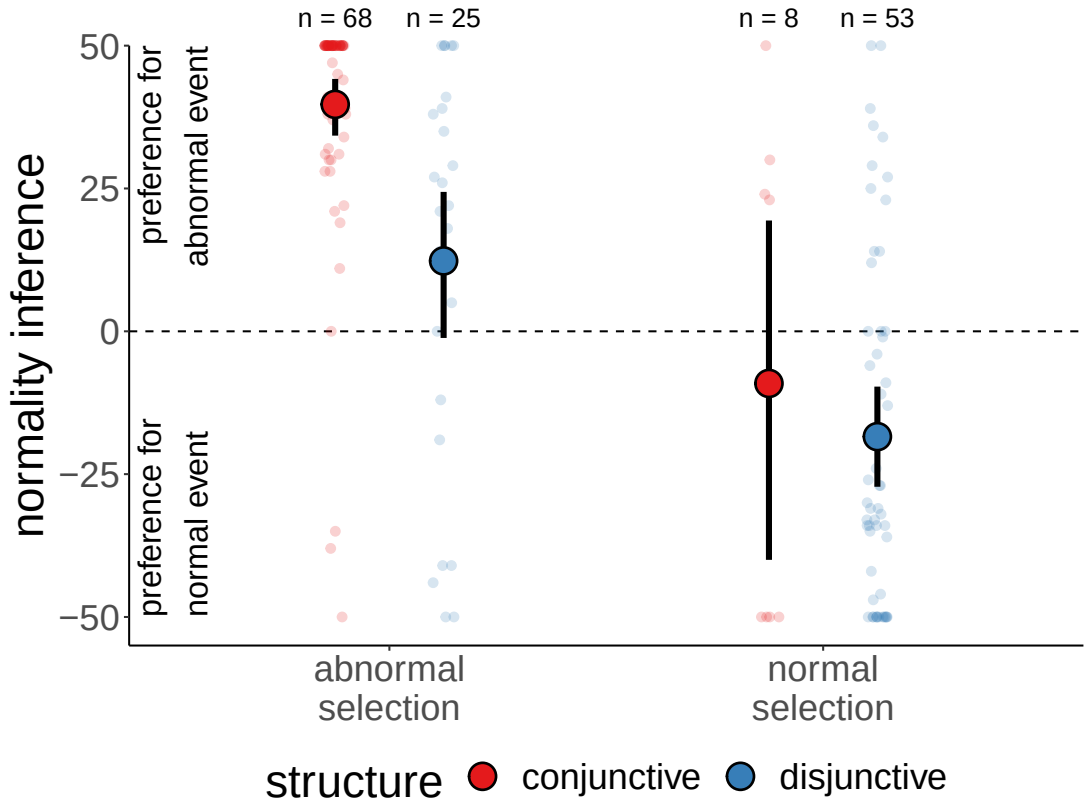

Figure 7

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

conjunctive structure (left-most data points in Figure 7) are more certain in their inference compared to participants who selected the normal cause for the disjunctive structure (right-most data points).

Discussion

The results of Experiment 1 are in line with previous literature on causal judgments (Gerstenberg & Icard, 2020; Icard et al., 2017; Kominsky et al., 2015), showing that the selection of causal explanations is affected both by normality and causal structure. People tend to select an explanation citing an abnormal cause in a conjunctive causal structure, but are more likely to select a normal cause when the structure is disjunctive ("Hypothesis 1"). Crucially, the experiment also confirmed our prediction of people's normality inferences from explanations ("Hypothesis 2"). People are more likely to infer that the cited event in the

explanation was abnormal when the underlying causal structure is conjunctive, compared to disjunctive.

In general, people's normality inference closely mirrored their own explanation preferences. People were more likely to infer that a cause was abnormal if they themselves selected an explanation citing an abnormal cause before. Interestingly, people's prior causal explanation not only influenced whether they inferred an abnormal or normal cause, but also the strength of their inference. The asymmetry in participants' causal selections reported previously (Icard et al., 2017; Kominsky & Phillips, 2019) also shows up in their inferences. In line with what has been found previously (cf. Gerstenberg & Icard, 2020), participants' tendency to select the abnormal cause in conjunctive structures is stronger than their preference to select the normal cause for disjunctive structures (cf. Figure 5). Correspondingly, participants were more certain that the cited cause in the explanation was abnormal in the conjunctive scenario, compared to how certain they were that the cited cause was normal in the disjunctive scenario.

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

Experiment 2: Inferring causal structure given normality

In Experiment 2, we test whether participants can infer the causal structure of a scenario based on whether a normal or abnormal event was cited in the explanation. In terms of our flatmate Suzy example, this means that the listener knows which of the two topics Suzy knows well (physiology) and which one she knows little about (anatomy). However, the listener doesn't know whether Suzy needs to pass both or just one of the tests in order to be accepted into medical school. When Suzy says "I got into med school because I passed anatomy.", we predict that the listener would infer that Suzy needed to pass both tests. This follows from the fact that people have a preference to cite abnormal causes in conjunctive structures, and normal causes in disjunctive structures. A listener can infer the causal structure by considering what they would have said in each of the two possible situations. Because they would be more likely to cite an abnormal cause in a conjunctive structure than in a disjunctive structure, they can infer the causal structure from the (ab)normality of the cause in the explanation (see Equation 1).

Methods

Participants and Design

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

& abnormal explanation ($N = 33$), *statistical normality & normal explanation* ($N = 27$), *prescriptive normality & abnormal explanation* ($N = 41$), and *prescriptive normality & normal explanation* ($N = 42$).In this experiment, participants were instructed about both causal structures, and the selection task was presented for both causal structures as well.

Statistical Normality: Selection Task

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

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

Statistical Normality: Inference Task

In the final inference task, participants received a diagram of a billiard ball scene in which both ball A and ball B went through the blocker and ball E went through the gate. However, the largest part of the billiard scene was grayed out (see Figure 3c). The causal diagram was missing the crucial information about where ball E was positioned at the beginning. Hence, the scene did not give away whether the causal structure was conjunctive or disjunctive.

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

Prescriptive Normality: Selection Task

Similar to Experiment 1, participants were instructed about the two employees Billy and Suzy. This time, however, they were informed that there are two offices in which Billy and Suzy sometimes work, depending on availability. The "Two-Door-Office" has a motion detector with a conjunctive structure, and the "One-Door-Office" has a motion detector with a disjunctive structure (see Figure 4). Given the confidentiality of the project, their boss sometimes instructs one of them to come into the office at 9am in the morning, while the other one is not allowed to come in that morning. In contrast to Experiment 1, normality was fixed: Who was allowed to come in and who not was *always* the same, independent of which office Suzy and Billy were currently working in. It was randomized across participants who of the two employees was supposed to come in, and who was supposed to stay away.

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

Prescriptive Normality: Inference Task

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

Results

Figure 5b shows participants' selections as a function of the causal structure and norm. Note that this time, we manipulated the causal structure within participants, so we asked each participant to indicate their selection for both causal structures. Table 4 shows the pattern of selections. Most participants ($n = 80$) selected the normal cause when the causal structure was disjunctive, and the abnormal cause when the structure was conjunctive. There was also a large group of participants ($n = 44$) who selected the abnormal cause for both structures.

Participants' selections were strongly affected by the causal structure. Overall, participants were more likely to select the abnormal cause for conjunctive causal structures (87%) compared to disjunctive structures (33%; see Table 5). There was also an interaction

between causal structure and norm. The difference in participants' selections between the conjunctive and disjunctive structures was stronger in the prescriptive norm condition compared to the statistical norm condition (see Figure 5b).

Figure 8 shows participants' inferences about the causal structure of the situation as a function of the type of explanation (citing an abnormal or a normal event) and the type of norm (statistical or prescriptive). Participants' inferences were affected by the normality of the explanation (see Table 6). Participants had a stronger preference to infer the conjunctive structure for explanations citing an abnormal event ($\text{Mean} = 74.07, \text{SD} = 30.26$) compared to a normal one ($\text{Mean} = 27.25, \text{SD} = 31.96$).

Note that unlike for the normality inferences $p(\text{normality} \mid \text{structure})$ in Experiment 1 (see Figure 6) which were asymmetrical around the midpoint of the scale, the structure inferences are symmetrical. As mentioned above, this follows directly from the application of Bayes' rule in order to determine $p(\text{structure} \mid \text{normality})$. Assuming that $p(\text{abnormal} \mid \text{conjunctive}) = 0.87$ and $p(\text{abnormal} \mid \text{disjunctive}) = 0.33$ (based on the probability with which participants actually selected the different causes), and assuming that the prior probability of a structure being conjunctive (or disjunctive) is

Table 4

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

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

Table 5

Experiment 2 – Causal selection*: Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian mixed effects regression model.* Note*: The units are log odds. The dependent variable (selection) was coded as 1 = abnormal cause, 0 = normal cause.*

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

term	estimate	lower 95% CI	upper 95% CI
intercept	−0.87	−1.48	−0.35
structure	−1.92	−2.77	−1.20
norm	−0.08	−0.54	0.42
structure:norm	0.46	0.02	0.91

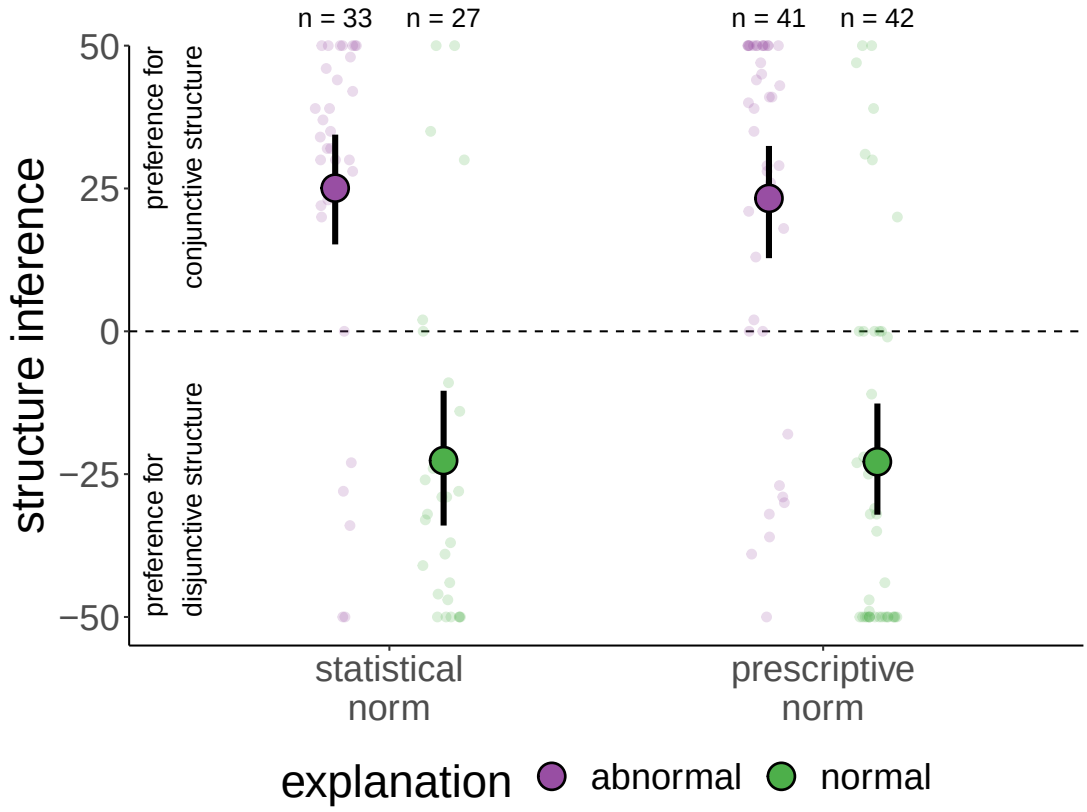

Figure 8

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

$p(\text{conjunctive}) = p(\text{disjunctive}) = 0.5$, it follows that

$$p(\text{structure} = \text{conjunctive} \mid \text{normality} = \text{abnormal}) =$$

$$\frac{p(\text{abnormal} \mid \text{conjunctive}) \cdot p(\text{conjunctive})}{p(\text{abnormal} \mid \text{conjunctive}) \cdot p(\text{conjunctive}) + p(\text{abnormal} \mid \text{disjunctive}) \cdot p(\text{disjunctive})} =$$

$$\frac{0.87 \cdot 0.5}{0.87 \cdot 0.5 + 0.33 \cdot 0.5} = 0.725. \tag{2}$$

As per the same logic, the probability of a conjunctive structure given that a normal cause was cited is $p(\text{structure} = \text{conjunctive} \mid \text{normality} = \text{normal}) = 0.275$. These predicted probabilities are symmetric around the midpoint and also correspond in magnitude very closely to participants' structure inferences (as expressed on a scale from −50 to 50 shown

in Figure 8).[8]

Individual differences

Figure 9 shows participants' structure inferences depending on what causal selections they themselves made. For example, the left-most data show the inference that participants made based on an abnormal explanation who themselves selected the abnormal cause both for the disjunctive (D) and conjunctive structure (C). What stands out is that the strength of the inference is strongest for the largest group of participants ($n = 80$) who selected the normal cause for the disjunctive structure and the abnormal cause for the conjunctive structure. For this group of participants, the difference between the group's mean inference based on an abnormal versus normal explanation is largest (Mean = 59.03). Even those participants who selected the abnormal event for both structures, or those who always selected the normal event, still had a preference for the conjunctive structure for abnormal explanations, and the disjunctive structure for normal explanations. Here, however, the difference between the inferences as a function of whether the explanation was abnormal or normal was weaker (Mean = 31.04 for the group of participants who always selected the abnormal cause, and Mean = 30.62 for the participants who always selected the normal cause).

We predict this pattern of inferences given the following two assumptions. 1) Participants on average have a stronger preference to cite the abnormal cause in conjunctive versus disjunctive situations. 2) The difference in preference is stronger for participants who selected the normal cause for disjunction, and the abnormal cause for conjunction, compared to participants who selected the abnormal (or normal) cause for both structures.

The first assumption is consistent with the fact that no matter what participants they themselves selected, they were more likely to infer the conjunctive struc-

Table 6

Experiment 2 – Structure inference*: Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian regression model. Note: For the dependent variable (structure rating), 100 = conjunctive and 0 = disjunctive.*

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

term	estimate	lower 95% CI	upper 95% CI
intercept	50.67	45.22	55.81
explanation	23.48	18.45	28.81
norm	0.56	−4.21	6.09
explanation:norm	0.48	−4.89	5.71

[8] We do not mean to imply that participants are explicitly computing Bayes' rule in order to infer the causal structure. We merely want to highlight that no additional theoretical machinery is required to account for the fact that while participants' normality inferences in Experiment 1 p(normality|structure) are asymmetric around the midpoint of the scale (see Figure 6), their structure inferences in Experiment 2 are symmetric (see Figure 8). The prediction that structure inferences in Experiment 2 should be symmetric follows from Bayes' rule (see Equation 2).

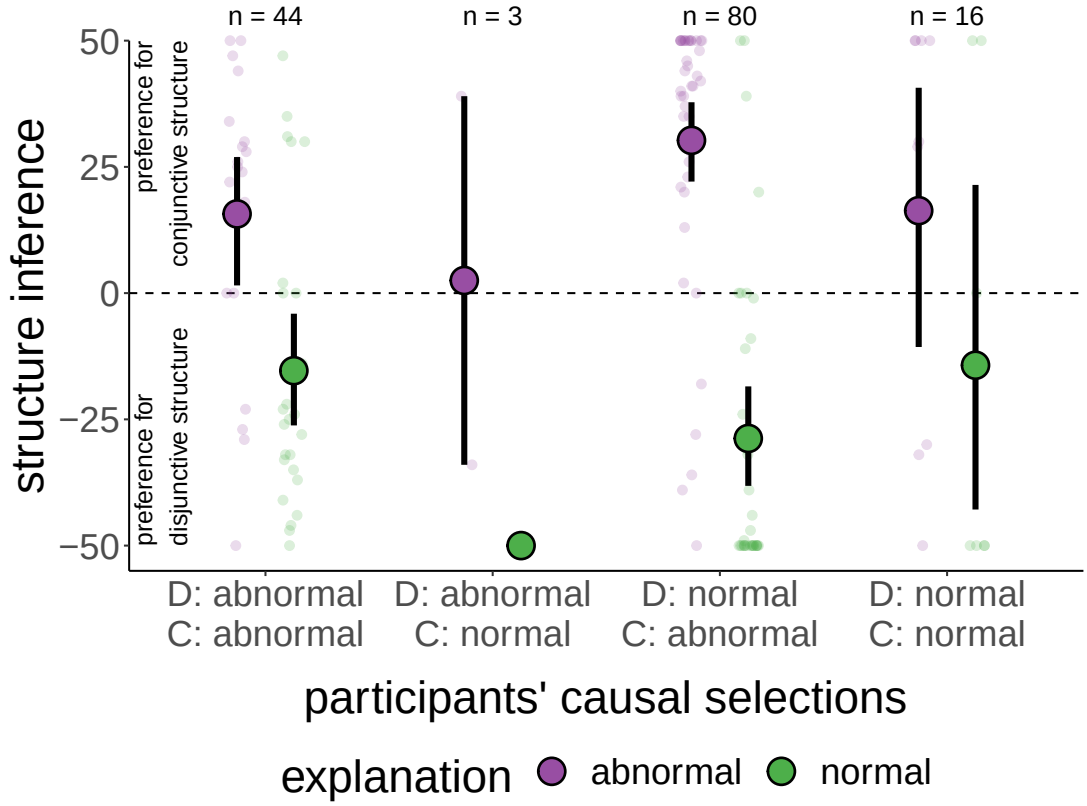

Figure 9

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

ture for abnormal explanations, and the disjunctive structure for normal explanations. The second assumption is consistent with the difference in the strength of the structure inferences between the groups. Specifically, we need to assume that the difference between $p(\text{abnormal} \mid \text{conjunctive})$ and $p(\text{abnormal} \mid \text{disjunctive})$ is greatest for the "D: normal, C: abnormal" group. For this group, we know from their choices that $p(\text{abnormal} \mid \text{conjunctive}) > 0.5$ and that $p(\text{abnormal} \mid \text{disjunctive}) < 0.5$. Let's assume that in fact for this group, $p(\text{abnormal} \mid \text{conjunctive}) = 0.8$ and $p(\text{abnormal} \mid \text{disjunctive}) = 0.3$. In this case, using Bayesian inference (cf. Equation 2), we would expect $p(\text{conjunctive} \mid \text{abnormal}) = 0.73$ and $p(\text{conjunctive} \mid \text{normal}) = 0.27$ (a difference of 0.46).

In contrast, for the "D: abnormal, C: abnormal" group, we know that both $p(\text{abnormal} \mid \text{conjunctive}) > 0.5$ and $p(\text{abnormal} \mid \text{disjunctive}) > 0.5$. For example, let's assume that for this group, on average, $p(\text{abnormal} \mid \text{conjunctive}) = 0.95$ and

$p(\text{abnormal} \mid \text{disjunctive}) = 0.6$. Here, we would expect $p(\text{conjunctive} \mid \text{abnormal}) = 0.61$ and $p(\text{conjunctive} \mid \text{normal}) = 0.39$ (a difference of 0.22). Note that the difference in the structure inference is weaker for the "D: abnormal, C: abnormal" group compared to the "D: normal, C: abnormal" group. The same rationale applies to the smaller "D: normal, C: normal" group of participants for which we know that both $p(\text{abnormal} \mid \text{conjunctive}) < 0.5$ and $p(\text{abnormal} \mid \text{disjunctive}) < 0.5$. So the difference in the strength of structure inferences between the groups of participants who differed in terms of which causes they themselves selected, is predicted by an application of Bayesian inference together with the two assumptions outlined above.

Discussion

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

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

The results of Experiment 2 show that the influence of norms and causal structure on causal explanations persists when people are able to directly compare and select explanations for both types of causal structures. People generally infer the causal structure from the normality of a cause in a way that tracks the interaction between normality and causal structure on causal explanations found in the literature (Gerstenberg & Icard, 2020; Icard et al., 2017; Kominsky et al., 2015). However, whether people themselves would give these explanations has an impact on the strength of their structure inferences. Participants who themselves selected causal explanations that matched the inference pattern (Abnormal: C, Normal: D) showed the strongest inference compared to those who selected deviating causal explanations. These results in particular shed new light on the crucial role of explanatory preferences for inferences.

General discussion

As Hilton (1990) observes, "the verb 'to explain' is a three-place predicate: *Someone* explains *something* to *someone*." (p. 65). Indeed, this communicative dimension is essential to our understanding of what explanation is and how it works. In this paper we have taken a first step in systematizing the inferential leaps that people make when comprehending explanations in explicitly communicative settings.

As a case study, we focused on the role of event normality and causal structure in explanations. In line with previous literature, we show that people prefer to explain an outcome in a conjunctive causal structure by referring to an abnormal cause, and in a in disjunctive causal structure, by referring to a normal cause. Crucially, we show that event normality and causal structure not only influence what explanations people give, but also what inferences people draw from others' explanations. When provided with a causal explanation about what happened and information about the causal structure, people are able to infer the normality of the cited cause. Even more strikingly, people can infer the causal structure of a scenario when provided with an explanation that cites a normal or abnormal cause. We show these inferential patterns for both statistical as well as prescriptive normality.

While these results may appear surprising, we have argued on theoretical grounds why this particular pattern of results should actually be expected. The results are consistent with the idea that a listener considers what they themselves would have said in the different possible situations, and interprets a speaker's explanation accordingly. By combining a wealth of previous research on causal judgment with established general principles of conversation, we begin to see how people can learn so much from simple statements such as "E because C." The present study opens up a productive method for systematizing the widely recognized but still poorly understood *compression* that is so characteristic of explanatory behavior (Hilton, 1990; Keil, 2006; Wilson & Keil, 1998).

In the present work we have used known patterns in causal judgment to derive predictions and shed light on the inferences people are able to draw from explanations. However, we may turn this back around and consider whether the communicative dimension of explanation may shed light on those very patterns that we have so far taken for granted. In the remainder of the paper we consider this broader issue.

A central assumption supporting our hypotheses is that people largely share intuitions about how norms and causal structure affect causal judgments. This raises two key follow-up questions:

1. Why do people share these intuitions in the first place?

2. Why do norms and causal structure affect causal explanations in the way that they do?

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

As mentioned earlier, there are already a number of proposals about 2. in the existing literature. It is thus worth considering how the communicative dimension of explanation studied here might interact with prominent accounts of the role of norms in causal judgment. We will consider three accounts that foreground different assumptions about what

mediates the effect of norms on causal explanations: 1) counterfactual reasoning, 2) blame and accountability, and 3) optimal interventions.

Counterfactual Reasoning

The *counterfactual reasoning* account (Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Knobe, 2009; Kominsky et al., 2015) draws on a substantial psychological link between causal explanation and *counterfactual relevance* (Byrne, 2016; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Kominsky & Phillips, 2019; Phillips, Luguri, & Knobe, 2015). According to this account, the meaning of a claim "E because C" involves counterfactual considerations, most notably the extent to which C was necessary for E, that is, the extent to which E would have occurred were it not for C (Lewis, 1973). Some counterfactual possibilities strike us as more relevant than others, and perhaps also come to mind more readily (Byrne, 2005, 2020; Johnson-Laird & Khemlani, 2017). Specifically, abnormal events tend to trigger counterfactual thoughts about what would have happened had things gone normally, while the reverse does not seem to hold (Kahneman & Miller, 1986). The relative availability of normal alternatives for abnormal causes makes these counterfactual necessity claims more easily verifiable, which in turn strengthens the relevant causal claim.

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

If the counterfactual reasoning account is correct, then norm effects and the interaction between norms and causal structure arise simply from the way our minds work, together with the basic semantics of "E because C" cashed out in terms of necessity and sufficiency. Thus, one possibility consistent with this type of account is that communicative efficiency is just a serendipitous byproduct of a more basic psychological pattern.

Blame and Accountability

A different line of research aims to explain the role of normality in causal selections in terms of blame and responsibility attributions. Some have argued that people's causal judgments are biased by a desire to assign blame to the abnormal factor (Alicke, 2000). According to this account, emphasizing the causal contribution of an abnormal cause allows people to validate their spontaneous blame response. Others argue that people's ordinary concept of causation is itself normative, with causal judgments being akin to judgments about responsibility (Sytsma, 2019; Sytsma & Livengood, 2019; Sytsma et al., 2012). Samland and Waldmann (2016) contend that these effects arise due to pragmatic factors in the context of norm violations and human agents. Rather than assessing an "actual causal" process, participants interpret the causal test question as a request to assign accountability (Samland & Waldmann, 2014, 2015, 2016). Accordingly, a speaker uses a causal explanation

"E because C" to communicate some form of attribution of responsibility or blame. On this type of account, we should therefore expect participants to make inferences consistent with the cited cause being blameworthy in some way.

While blame-oriented accounts explicitly address the communicative function of causal explanations, and provide a plausible explanation for causal selections in case of prescriptive norms and human agent causation, it is less clear how they would work for inanimate objects and statistical normality. When provided with the explanation "Ball E went through the gate because Ball A went through the blocker", it seems questionable whether the recipient will interpret this statement as an expression of blame or responsibility attribution centered on the ball. Moreover, the effect of normality has been shown for outcomes that are positive, neutral and bad in nature (Icard et al., 2017; Reuter, Kirfel, Van Riel, & Barlassina, 2014). In addition, these accounts leave open why in a disjunctive structure, people blame the agent that adheres to the prescriptive norm. Some have argued that prior expectations or the agent's foreseeability of the outcome might have an impact on causal judgments in disjunctive structures (Kirfel & Lagnado, 2018, 2019). However, at the current theoretical state, these blame-oriented models do not explain how a normative judgment is made in circumstances other than a clear rule violation, or when an action results in a bad outcome (Alicke & Rose, 2012; Samland & Waldmann, 2016). Without a more precise account of responsibility or blame, it doesn't seem possible to identify a sensible and unequivocal blame response from the diverse range of causal explanations that are impacted by normality. This makes it difficult for accounts referring to blame or accountability to predict the overall pattern found in our experiments, even the basic patterns for the selection tasks.

Explanations point out optimal interventions

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

How might an optimal intervention account account for the results in this paper? Specifically, in what sense might it be better to intervene on an abnormal cause in con-

junctive structures, and a normal cause in disjunctive structures? Suppose that an optimal intervention is one that makes the largest difference to the probability of the outcome of interest. Consider our billiard ball setting with a conjunctive structure: Ball A and ball B both need to pass their blockers in order for ball E to go through the gate. Suppose ball A has a 20% chance of going through the blocker, while ball B has an 80% chance. We want to compare $p(E|do(C)) - p(E|do(\neg C))$, where C is either ball A or ball B going through the block.[9] For A we get $p(E|\text{do}(A)) - p(E|\text{do}(\neg A)) = 0.8 - 0 = 0.8$. There is an 80% chance that ball E will go through the gate if we make sure that ball A goes through the blocker, and a 0% chance if we prevent ball A from going through the blocker. In contrast, for ball B we get $p(E|\text{do}(B)) - p(E|\text{do}(\neg B)) = 0.2 - 0 = 0.2$. Thus, in a conjunctive scenario intervening on the less likely event makes the biggest difference to the probability of the outcome. To make the biggest difference to the outcome, a person should intervene on ball A rather than on ball B.

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

This simple analysis is suggestive of a more general formalization of the optimal intervention account. However, there is a potential tension between this story and the communicative role of explanation revealed in our experiments. As we saw, people are able to infer missing probabilistic, normative, or causal information from an explanation. Given a full understanding of the causal setup – and evidently participants largely achieved such understanding in our experiments – any downstream decision-making could simply make appropriate use of this knowledge to calculate how desirable any given action would be. With full knowledge of the situation, what need is there for highlighting a variable that a speaker would deem especially worthy of consideration?

There are several ways of resolving this puzzle. First, in realistic scenarios we cannot always expect that listeners (or speakers for that matter) have full knowledge of the situation. For example, to make the scenario only slightly more challenging, suppose that our listener knows neither the causal structure nor the normative status of variables. From an utterance "E went through the gate because A went through its blocker," such a person could at best infer that, either the causal structure is conjunctive and A was unlikely, or the structure is disjunctive and A is likely. Importantly, it may not matter which state of affairs obtains – in many scenarios it will just be critical that the listener knows which variable to manipulate to make the largest difference to the outcome.

A second relevant consideration is that computing the best course of action may simply be too complex. After all, the listener needs to go through two steps: first update their model of the world appropriately, and second compute the best option (e.g., the one expected to increase the probability of the desired outcome most). In case a speaker has already determined what they expect the best future intervention to be, they can communicate this directly, bypassing costly computation on the part of the listener.

The communicative framework sketched in the introduction focused on the resolu-

[9] Here E stands for the event of ball E going through the gate. The do() operator indicates that we fix the event via an intervention (making it either true or false, see Pearl, 2000). However, in the simple case here these expressions are equal to the respective conditional probabilities, $p(E|C)$ and $p(E|\neg C)$.

tion of uncertainty about a situation, and our experiments similarly highlight this epistemic aspect of linguistic interpretation. However, it may be that the ultimate explanation for the specific patterns of norm effects we see on causal judgments is properly framed in a broader communicative story. Indeed, on a more general analysis of *signals*, appropriate for a much wider array of communicative situations, coordination is assumed to center around the *receiver* choosing the right action as a function of the *sender's* chosen signal (Lewis, 1969; Skyrms, 2010). Making an appropriate inference about the world is just one specific type of action that might be relevant, and in any case it will often merely be a means to some more practical end.

This framework also offers potential insight into the asymmetric way in which norms and causal structure affect people's causal selections. Recall that participants are more likely to select the abnormal cause in conjunctive structures than they are to select the normal cause in disjunctive structures (see Figure 5), and that this asymmetry in how explanations are generated is reflected in the inferences that people draw (see Figure 6). How may this pattern of causal selections arise from communicative pressures?

Suppose that causal judgments are sensitive to other communicative pressures aside from those discussed above. In particular, it seems reasonable to assume that identifying an abnormal event will often be helpful, especially when the listener is unaware of it. After all, when an alternative, *normal* event can reasonably be assumed, mentioning the abnormal event will be strictly more informative. We thus have (at least) two communicative pressures that may shape causal explanations: being generally informative and highlighting a variable that would be a good point of intervention. In conjunctive structures these two pressures both focus attention on the abnormal event. However, in disjunctive structures they pull in different directions. These conflicting pressures may account for the asymmetric pattern observed in the data.[10]

Of course, as has been long appreciated in the literature (see, e.g., Coffa, 1974), a purely forward-looking approach to causal explanation, focusing only on a variable's future causal potential, may flounder on cases where backward-looking and forward-looking dimensions come apart. If a person takes a treatment for an illness but then gets better independently, the treatment should not be part of the explanation no matter how effective it is in general. What matters is not how good of an intervention we expect taking the treatment to be in the future, but rather the fact that *in this particular instance* the treatment did not cause the improvement. Negotiating the balance between these two dimensions of causal explanation is an important challenge for any adequate theory of explanation (cf. Hitchcock, 2017).

Toward a process account of inference from explanation

The theoretical proposal motivating the present work was intended to be reasonably neutral with respect to the different approaches to causal judgment just canvassed. The

[10] It also matters how people construe the notion of an optimal intervention. For example, it's possible that people think an optimal intervention is the one that is most likely to make the outcome happen (optimal intervention = $max(p(E|do(C)))$, rather than the one that makes the biggest difference to the outcome (optimal intervention = $max(p(E|do(C)) - p(E|do(\neg C)))$. In that case, for conjunctive structures, one should intervene on the abnormal event, whereas for disjunctive structures it doesn't matter since either event is sufficient to make the outcome happen.

starting point for our four hypotheses was not so much a theoretical commitment to any one of these approaches as a set of empirical patterns that such accounts would aim to explain. Our hypothesis was that, whatever the ultimate source of the patterns, people would draw upon their own causal intuitions to make systematic inferences from others' explanations, consistent with very general principles of ordinary conversation (Hypotheses 2, 3, and 4). What might our experiments validating these hypotheses tell us about the actual process by which people draw such inferences?

Probabilistic theories of conversational pragmatics typically aim at computational-level analyses, characterizing at a high level the principles facilitating successful communication over experimental populations (e.g., Franke & Jäger, 2016; Goodman & Frank, 2016), and our proposal can also be understood at this level. At the same time, just as these frameworks can be used to investigate more detailed individual-level questions (e.g., Franke & Degen, 2016), the present studies also shed light on such questions. Our Hypothesis 4 proposed not just that people's inferences could be predicted by the population-level trends (as in Hypotheses 2 and 3), but that they would be correlated with their own individual judgments. We saw this in both experiments. One potential explanation of this finding is that people are consulting their own intuitions about what would be reasonable to say in the two alternative circumstances (e.g., depending on whether or not Suzy was supposed to come into the office, as in Experiment 1).

Ultimately, however, a more comprehensive process-level account cannot remain neutral on the question of what drives the causal intuitions in the first place. Indeed, as discussed above, the different proposals (counterfactual, blame/accountability, optimal intervention) fit into the broader communication-theoretic account in different ways. Most saliently, unlike counterfactual-centered accounts, blame/accountability and optimal intervention accounts both tend to put communication first: it is a communicative goal that itself explains the patterns we see with normality and causal structure. While this difference may have less of an impact on how we understand a listener (our focus in this paper), it matters a great deal in how we understand what a speaker intends to convey with a causal explanation.

Future directions

In this paper, we focused on people's inferences about event normality and causal structure. Future work will examine to what extent our communication-theoretic account can capture a broader array of inference patterns. For example, research has shown that people often prefer to cite early or late events as the cause of the outcome in a way that is sensitive to causal structure (Brickman, Ryan, & Wortman, 1975; Gerstenberg & Lagnado, 2012; Glautier, 2008; Hilton, McClure, & Sutton, 2010; McClure, Hilton, & Sutton, 2007; Spellman, 1997). Given this systematic pattern, we would predict that people should be able to infer the temporal order of events from explanations given knowledge about the causal structure (Henne et al., 2021). As another illustrative case study, we are interested in exploring the kinds of inferences that people draw from social evaluations. In the social domain concepts like *blame* are closely related to explanation (see, e.g. Gerstenberg et al., 2018; Malle, 2021). For example, if a goalkeeper is blamed for a soccer team's loss, this suggests a different inference about what happened (e.g., the goalkeeper didn't block an

easy shot) compared to a situation in which the striker was blamed (e.g., the striker missed an easy goal).

Moving beyond explanation, the patterns studied in our experiment are pervasive in human cognition more generally. For example, research on conditional reasoning has shown that people's assumptions about the causal structure (Bonnefond, Kaliuzhna, Van der Henst, & De Neys, 2014; Byrne, 1989; Byrne, Espino, & Santamaria, 1999; Espino & Byrne, 2020) and their normative expectations about the frequency of events (Oaksford & Chater, 1994, 2003) affect what inferences people draw. Precisely what people infer from conditional statements is still very much under investigation (Barrouillet, Gauffroy, & Lecas, 2008; Collins, Krzyżanowska, Hartmann, Wheeler, & Hahn, 2020; Khemlani & Johnson-Laird, 2013; Sebben & Ullrich, 2021; Skovgaard-Olsen, Stephan, & Waldmann, 2021). Given the tight relationship between conditionals and causality (e.g. Goldvarg & Johnson-Laird, 2001; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007), we suspect that inferences from conditional statements, just like inferences from explanations, may be illuminated by considering what role these statements play in communication (see also Evans, 2005; Johnson-Laird & Byrne, 2002; Sebben & Ullrich, 2021).

Conclusion

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

Acknowledgments

We thank Jonathan Kominsky, Jonathan Phillips, Joshua Knobe, Nadya Vasilyeva, Ruth Byrne, and the Australasian Experimental Philosophy Group for providing feedback on the project and comments on the paper. We are grateful to the anonymous reviewers for their constructive feedback. We also thank the members of the Causality in Cognition Lab at Stanford University for feedback and discussion. Part of this research was presented at the 42nd Cognitive Science Conference in 2020, and at the Society for Philosophy and Psychology Conference in 2021.

References

Achinstein, P. (1983). *The nature of explanation.* Oxford University Press.

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

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

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

Barrouillet, P., Gauffroy, C., & Lecas, J.-F. (2008). Mental models and the suppositional account of conditionals. *Psychological review*, *115*(3), 760.

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

Bonnefond, M., Kaliuzhna, M., Van der Henst, J.-B., & De Neys, W. (2014). Disabling conditional inferences: an eeg study. *Neuropsychologia*, *56*, 255–262.

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

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

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83.

Byrne, R. M. (2005). *The rational imagination: How people create alternatives to reality.* MIT Press.

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

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, ijcai-19* (pp. 6276–6282).

Byrne, R. M. (2020). The counterfactual imagination: The impact of alternatives to reality on morality. In A. Abraham (Ed.), *The cambridge handbook of the imagination* (pp. 529–547). Cambridge University Press.

Byrne, R. M., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language*, *40*(3), 347–373.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

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

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

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

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

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

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

Collins, P. J., Krzyżanowska, K., Hartmann, S., Wheeler, G., & Hahn, U. (2020). Conditionals and testimony. *Cognitive Psychology*, *122*, 101329.

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

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

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

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

Espino, O., & Byrne, R. M. (2020). The suppression of inferences from counterfactual conditionals. *Cognitive science*, *44*(4), e12827.

Evans, J. S. B. (2005). The social and communicative function of conditional statements. *Mind & Society*, *4*(1), 97–113.

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

Franke, M., & Degen, J. (2016). Reasoning in reference games: individual- vs. population-level probabilistic modeling. *PLoS ONE*, *11*(5), 1-25.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft 2016*, *35*(1), 3-44.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5-19.

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

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

Gerstenberg, T., & Icard, T. F. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

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

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

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

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Cognition*, *23*(5), 389-407.

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

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive science*, *25*(4), 565–610.

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

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

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

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

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

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

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, *66*, 413–457.

Hanna, E., & Meltzoff, A. N. (1993). Peer imitation by toddlers in laboratory, home, and day-care contexts: Implications for social learning and memory. *Developmental psychology*, *29*(4), 701.

Harinen, T. (2017, jul). Normal causes for normal effects: Reinvigorating the correspondence hypothesis about judgments of actual causation. *Erkenntnis*. Retrieved from `https://doi.org/10.1007%2Fs10670-017-9876-4` doi: 10.1007/s10670-017-9876-4

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

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

Hemmatian, B., & Sloman, S. A. (2018). Community appeal: Explanation without information. *Journal of Experimental Psychology: General*, *147*(11), 1677.

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

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

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Brighton, UK: Harvester Press.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes. *European Journal of Social Psychology*, *40*(3), 383–400.

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

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

Hitchcock, C. (2017). Actual causation: What's the use? In *Making a difference: Essays on the philosophy of causation*. Oxford University Press.

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

Horwich, P. (1998). *Meaning*. Oxford University Press.

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

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 785. Retrieved from `https://doi.org/10.1016%2Fj.tics.2016.08.007` doi: 10.1016/j.tics.2016.08.007

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological review*, *109*(4), 646.

Johnson-Laird, P. N., & Khemlani, S. (2017). Mental models and causation. *Oxford handbook of causal reasoning*, 1–42.

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

Kamide, Y. (2012, July). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71. doi: 10.1016/j.cognition.2012.03.001

Keil, F. (2006). Explanation and understanding. *Annual review of psychology*, *57*, 227.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4–20.

Kirfel, L., & Lagnado, D. A. (2018). Statistical norm effects in causal cognition. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 615–620). Austin, TX: Cognitive Science Society.

Kirfel, L., & Lagnado, D. A. (2019). I know what you did last summer (and how often). epistemic states and statistical normality in causal judgments. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

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

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: intuition and diversity* (Vol. 2). The MIT Press.

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.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. *Causal learning: Psychology, philosophy, and computation*, 154–172.

Levinson, S. C. (2000). *Presumptive meanings*. MIT Press.

Lewis, D. (1969). *Convention: A philosophical study*. Wiley.

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

Lewis, D. (1986). Causal explanation. *Philosophical Papers*, *2*, 214–240.

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. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

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.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014, Apr). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186. Retrieved from `http://dx.doi.org/10.1080/1047840x.2014.877340` doi: 10.1080/1047840x.2014.877340

Marcus, G., & Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Pantheon.

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.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. *PsyArXiv*. Retrieved from `https://psyarxiv.com/nq53z`

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind & Language*, *18*(4), 359–379.

Over, D. E., Hadjichristidis, C., Evans, J. S. B., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62-97.

Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, *121*(1-2), 93–149.

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

Phillips, J., & Cushman, F. (2017, apr). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*(18), 4649–4654. Retrieved from `https://doi.org/10.1073%2Fpnas.1619717114` doi: 10.1073/pnas.1619717114

Phillips, J., Luguri, J., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Potochnik, A. (2016, Dec). Scientific explanation: Putting communication first. *Philosophy of Science*, *83*(5), 721–732. Retrieved from `http://dx.doi.org/10.1086/687858` doi: 10.1086/687858

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Reuter, K., Kirfel, L., Van Riel, R., & Barlassina, L. (2014). The good, the bad, and the timely: how temporal order and moral judgment influence causal selection. *Frontiers in psychology*, *5*, 1336.

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

Samland, J., & Waldmann, M. R. (2014). Do social norms influence causal inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2015). Highlighting the causal meaning of causal test questions in contexts of norm violations. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2092–2097). Austin, TX: Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176. Retrieved from `https://doi.org/10.1016%2Fj.cognition.2016.07.007` doi: 10.1016/j.cognition.2016.07.007

Schuster, S., & Degen, J. (2019). Speaker-specific adaptation to variable use of uncertainty expressions. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 7). Cognitive Science Society.

Sebben, S., & Ullrich, J. (2021). Can conditionals explain explanations? a modus ponens model of b because a. *Cognition*, *215*, 104812.

Skovgaard-Olsen, N., Stephan, S., & Waldmann, M. (2021). Conditionals and the hierarchy of causal queries. *Journal of Experimental Psychology: General*, *1*.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

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

Strevens, M. (2008). *Depth*. Harvard University Press.

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

Sytsma, J., & Livengood, J. (2019). Causal attributions and the trolley problem.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.

Turnbull, W., & Slugoski, B. R. (1988). Conversational and linguistic processes in causal attribu-

tion. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 66–93). Brighton, UK: Harvester Press.

van Fraassen, B. (1980). *The scientific image*. Oxford University Press.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018, April). Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*. doi: 10.1111/cogs.12605

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58.

Whiten, A. (2002). Imitation of sequential and hierarchical structure in action: Experimental studies with children and chimpanzees.

Wilson, R. A., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and machines*, *8*(1), 137–159.

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

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

Woodward, J. (2011). Psychological studies of causal and counterfactual reasoning. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*. Oxford: Oxford University Press.

Woodward, J. (2014). *A functional account of causation.* Retrieved from `http://philsci-archive.pitt.edu/10978/`

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016, April). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. doi: 10.1016/j.jml.2015.08.003

Appendix

Frequentist analysis of the experiment results

In Experiment 1, there was a significant effect of structure on participants' causal selections, $\chi^2 = 58.77, p < .001,$ Cramer's $V = 0.62$. There was no significant effect of norm, and no interaction effect between norm and structure.

Participants' inferences about the normality of the cause were significantly affected by structure $F(1, 150) = 66.59, p < .001, \eta_p^2 = 0.32$. There was no significant effect of norm, and no interaction effect between norm and structure.

In Experiment 2, there was a significant effect of structure on participants' causal selections, $\chi^2 = 26.44, p < .001,$ Cramer's $V = 0.30$. There was no significant effect of norm. There was a significant interaction effect between structure and norm $\chi^2 = 4.19, p = .04,$ Cramer's $V = 0.12$.

Participants' inferences about the causal structure of the situation were significantly affected by whether the explanation cited a normal or abnormal cause $F(1, 139) = 77.15, p < .001, \eta_p^2 = 0.36$. There was no significant effect of norm, and no interaction effect between norm and explanation.