Cognition 223 (2022) 105036

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Confidence and gradation in causal judgment

Kevin O'Neill [a,b,h,*], Paul Henne [f,g], Paul Bello [h], John Pearson [a,b,d,e], Felipe De Brigard [a,b,c]

[a] Center for Cognitive Neuroscience, Duke University, United States of America
[b] Department of Psychology and Neuroscience, Duke University, United States of America
[c] Department of Philosophy, Duke University, United States of America
[d] Department of Biostatistics & Bioinformatics, Duke University, United States of America
[e] Department of Electrical and Computer Engineering, Duke University, United States of America
[f] Department of Philosophy, Lake Forest College, United States of America
[g] Neuroscience Program, Lake Forest College, United States of America
[h] Navy Center for Applied Research in Artificial Intelligence, U.S. Naval Research Laboratory, United States of America

ARTICLE INFO

Keywords:
Causation
Causal judgment
Norms
Gradation
Confidence

ABSTRACT

When comparing the roles of the lightning strike and the dry climate in causing the forest fire, one might think that the lightning strike is more of a cause than the dry climate, or one might think that the lightning strike completely caused the fire while the dry conditions did not cause it at all. Psychologists and philosophers have long debated whether such causal judgments are graded; that is, whether people treat some causes as stronger than others. To address this debate, we first reanalyzed data from four recent studies. We found that causal judgments were actually multimodal: although most causal judgments made on a continuous scale were categorical, there was also some gradation. We then tested two competing explanations for this gradation: the *confidence explanation*, which states that people make graded causal judgments because they have varying degrees of belief in causal relations, and the *strength explanation*, which states that people make graded causal judgments because they believe that causation itself is graded. Experiment 1 tested the confidence explanation and showed that gradation in causal judgments was indeed moderated by confidence: people tended to make graded causal judgments when they were unconfident, but they tended to make more categorical causal judgments when they were confident. Experiment 2 tested the causal strength explanation and showed that although confidence still explained variation in causal judgments, it did not explain away the effects of normality, causal structure, or the number of candidate causes. Overall, we found that causal judgments were multimodal and that people make graded judgments both when they think a cause is weak and when they are uncertain about its causal role.

1. Introduction

Lightning struck a tree in a forest, and a forest fire ensued. When asked what caused the fire, most people would be inclined to point to the lightning strike, since the fire would not have ensued without the strike of lightning. But many other factors, including the presence of oxygen, the location of the tree, and the lack of rain, are also necessary for the fire to happen. So, why do people tend to identify the lightning strike, and not—for instance—the lack of rain, as the cause of the fire (Danks, 2017; Halpern & Hitchcock, 2015; Icard, Kominsky, & Knobe, 2017; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Morris, Phillips, Gerstenberg, & Cushman, 2019)? More generally, why do people judge certain events as more or less causally relevant than others?

A vast amount of research on causal judgment has sought to answer this question. A popular explanation is that lightning is statistically abnormal, which allows it to stand out from more normal events like the lack of rain (Gerstenberg & Icard, 2020; Hart & Honoré, 1985; Henne, O'Neill, Bello, Khemlani, & De Brigard, 2021; Henne, Pinillos, & De Brigard, 2017; Hilton & Slugoski, 1986; Icard et al., 2017; Kahneman & Miller, 1986; Knobe & Fraser, 2008; McGrath, 2005). In addition to normality, research also indicates that people are more likely to judge events as causal when they are temporally recent (Henne, Kulesza, Perez, & Houcek, 2021; Lagnado & Channon, 2008; Spellman, 1997), necessary or sufficient (Icard et al., 2017; Pearl, 2009), robust to a range of background circumstances (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Hitchcock, 2012; Lombrozo, 2010; Quillien, 2020; Vasilyeva, Blanchard, & Lombrozo, 2018; Woodward, 2006), intentional or

* Corresponding author at: Department of Psychology & Neuroscience, Duke University, 417 Chapel Drive, Durham, NC 27708, United States of America.
E-mail address: kevin.oneill@duke.edu (K. O'Neill).

https://doi.org/10.1016/j.cognition.2022.105036
Received 18 December 2020; Received in revised form 20 December 2021; Accepted 17 January 2022
Available online 29 January 2022
0010-0277/Published by Elsevier B.V.

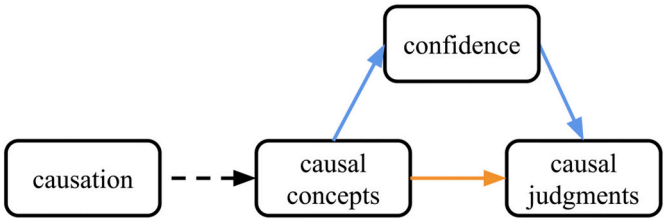

Fig. 1. Two explanations of gradation in causal judgments.
People's concept(s) of causation (i.e., what they think causation is) may or may not depend on what causation actually is (dashed arrow). Under the *causal strength explanation* (orange arrow), people's causal judgments depend directly on their concept(s) of causation. So, if causal judgments are graded, it is because people conceive of causation as being graded. Under the *confidence explanation* (blue arrows), people's causal judgments depend on their degree of belief in a causal relation, which in turn depends on their concept(s) of causation. On this explanation, if causal judgments are graded, it is because people have varying degrees of belief in a (graded or non-graded) causal relation. These two explanations are competing but not mutually exclusive: gradation in causal judgments could be partly due to gradation in people's concept of causation and partly due to gradation in their belief about a causal relation.

agentive (Alicke, Rose, & Bloom, 2011; Kirfel & Lagnado, 2021; Lagnado & Channon, 2008), connected through a physical process (Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010), and when there are few alternate causes (Lagnado, Gerstenberg, & Zultan, 2013).

There are at least two interpretations of what it means for people to judge some events as more causally relevant than others. One natural possibility is that people make causal judgments according to their representations of *causal strength*: that is, they think that causation is graded and that the lightning is more of a cause of the fire than the lack of rain. Another underexplored possibility is that people make their judgments according to their degree of *confidence*: they simply mean to indicate that they are more certain in their belief that the lightning caused the fire than they are that the lack of rain caused it. The primary goal of the current paper is to help to adjudicate between these two broad interpretations of effects on causal judgments—one relying on causal strength and another relying on confidence—in order to determine whether people treat causation as graded. To further clarify these two interpretations, consider the different components that are brought to bear whenever we make a causal judgment, as illustrated in Fig. 1. As we discuss next in further detail, since some of these components may or may not be graded, explanations as to why causal judgments admit degrees may also vary. Then, with this conceptual framework in mind, we return to the above interpretations of causal judgments and derive the empirically testable predictions made by each.

1.1. Layers of gradation in causal judgment

1.1.1. Causation

We use the term 'causation' to refer to what causation *actually* is, regardless of what people *think* causation is. Philosophers still actively debate whether causation is graded. While early accounts (e.g., Hume, 1748/2000; Lewis, 1973) endorsed non-graded definitions of causation, the dominant view is now that causation does come in degrees (Danks, 2013; Danks, 2017; Halpern & Hitchcock, 2015; Hart & Honoré, 1985; Hitchcock & Knobe, 2009; Sprenger, 2018) or are at least consistent with this possibility (Dowe, 1992; Salmon, 1994). Nevertheless, some philosophers still argue that degrees of causation are illusory and that causation is not graded (Bernstein, 2017; Kaiserman, 2016; Kaiserman, 2018; Sartorio, 2020). Since we are interested in non-philosophers' judgments of causation and since it remains to be shown whether people's causal judgments depend at all on what causation actually is (indicated by the dashed arrow in Fig. 1), we will not take a stance on whether causation itself is graded.

1.1.2. Concepts of causation

In contrast to causation, *concepts of causation* refer to what people think causation is, or how people mentally represent causation. Psychologists typically distinguish between three broad accounts of causal concepts. Dependence accounts propose that people think of an event as a cause when they believe the cause made a difference to the effect, or when the effect depends on the cause in some way (Cheng, 1997; Cheng & Novick, 1990; Gerstenberg et al., 2021; Icard et al., 2017; Morris et al., 2018; Pearl, 2009; Quillien, 2020; Spellman, 1997; Spellman & Ndiaye, 2007). Production accounts propose that people think causes transfer forces to their effects (Wolff, 2007; Wolff et al., 2010). Finally, causal pluralist accounts hold that people combine or contextually switch between multiple concepts of causation (Godfrey-Smith, 2009; Hall, 2004; Lombrozo, 2010).

For our purposes, it does not matter which of these concepts people actually use; all we are interested in is whether their concept is graded. Specifically, if people have a graded concept of causation, they will represent causation on a continuum of causal strength from non-causal to totally causal, and they will treat values along this continuum as directly comparable (e.g., the lightning is more of a cause of the fire than the lack of rain). If people have a non-graded concept of causation, on the other hand, they will think that an event is either causal or not, and they will disagree that any one cause is stronger than any other (e.g., both the lightning and the lack of rain caused the fire). Most researchers have come to endorse the position that people conceive of causation as being graded (e.g., Cheng, 1997; Gerstenberg et al., 2021; Icard et al., 2017; Jenkins & Ward, 1965; Lombrozo, 2007; Lombrozo, 2010; Morris et al., 2018; Quillien, 2020; Spellman, 1997). But many extant theories do not posit that people have graded concepts of causation, instead relying on qualitative distinctions between different types of causal verbs (e.g., 'caused' vs. 'allowed' vs. 'prevented'; Bello & Khemlani, 2015; Bello, Lovett, Briggs, & O'Neill, 2018; Goldvarg & Johnson-Laird, 2001; Khemlani, Barbey, & Johnson-Laird, 2014; Khemlani, Wasylyshyn, Briggs, & Bello, 2018; Wolff, 2007; Wolff et al., 2010). In any case, if people have a graded concept of causation, we should find that distributions of causal judgments are also graded, and that at least some of the gradation in causal judgments directly reflects participants' representations of causal strength.

1.1.3. Causal judgment

A causal judgment is a behavioral report of some belief in a causal relation that is measured in psychological experiments. Usually, this amounts to a response to the question "To what extent do you agree with the statement 'X caused Y?'" or "To what extent do you agree with the statement 'Y because X?'" on a Likert scale ranging from "strongly disagree" to "strongly agree" (e.g., Icard et al., 2017; Kominsky et al., 2015; Henne, Bello, Khemlani, & De Brigard, 2019). But researchers have also used other scales that ask about causal responsibility (e.g., "To what degree is X responsible for Y?"; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014; Lagnado et al., 2013) or that ask directly about degrees of causation (e.g., "To what degree did X cause/prevent Y?"; Gerstenberg et al., 2021; Lagnado & Channon, 2008).

As suggested by the case of the lightning and the forest fire, it is widely known that mean causal judgment varies continuously across the range of the scale under different conditions (e.g., Icard et al., 2017; Morris et al., 2019). However, researchers have focused less attention on whether distributions of individual causal judgments are in fact graded. There is a trivial sense in which individual judgments must be graded due to measurement error alone, since judgments are unlikely to have perfect test-retest reliability. In addition to between-participant variability in causal judgments, there is also substantial variability within participants, since participants might attend to different aspects of the stimulus, use a different strategy to make a causal judgment, or provide slightly different ratings in response to the same stimulus (Kolvoort, Davis, van Maanen, & Rehder, 2021). But no one has addressed the extent to which graded causal judgments reflect signal or noise. If most causal judgments are clustered around a graded mean, we can infer that there is between-participant agreement on this judgment and treat the gradation as

signal. But there are at least two other distributions of causal judgments that could also lead to a graded mean with different interpretations. If causal judgments are categorical even on a continuous scale (i.e., they lie mostly at the scale extremes with few graded judgments), we can infer that participants employ a non-graded concept of causation and that any gradation in causal judgments beyond this categorical distinction reflects noise rather than signal. In this case, we would be better off measuring causal judgments on a categorical scale to reduce such noise. If causal judgments are often graded but are widely dispersed or uniformly distributed, we can infer that even though participants do tend to make fine distinctions in the gradation of a cause, they disagree with each other to a large extent. Unless this large amount of between-participant variability can be meaningfully explained away, we may be better off treating it as noise. Clearly, different distributions of causal judgments can lead to an average that appears graded. So, before asking why causal judgments are graded, we first need to ask whether they are in fact graded at all.

1.1.4. Confidence

Finally, in addition to assessing whether and to what degree an event is a cause, it is also possible that people can be more or less confident about those assessments. We will use the term 'confidence' to refer to the degree to which one believes in a causal relation irrespective of whether the content of that belief invokes a graded or non-graded concept of causation. While the question of whether belief is graded is a question for epistemology (Jackson, 2020), there is some evidence that people report varying levels of belief in causal relations. Though only a handful of studies on general causal judgments—judgments of whether a type of event generally causes a type of outcome—have investigated confidence in causal judgments, these studies have shown that people report being more confident in their causal judgments when they have more relevant data (Collins & Shanks, 2006; Liljeholm & Cheng, 2009; Perales & Shanks, 2003; Schlottmann & Anderson, 1993; Shanks, 1987; Shou & Smithson, 2015). As such, we will take it for granted that people can have varying degrees of belief or confidence in their causal judgments.

1.2. Two explanations of gradation in causal judgment

With this theoretical framework in mind, we can return to the question of what participants communicate when they judge one event to be more causally relevant to an outcome than another event. Under the standard interpretation, which we call the *causal strength explanation*, causal judgments are graded because people have graded concepts of causation (Icard et al., 2017; Morris et al., 2019; orange path in Fig. 1). An immediate prediction of this account is that at least some of the gradation in causal judgments (if there is any) should be directly attributable to participants' graded representations of causal strength. For example, people should give higher causal judgments of the lightning than of the lack of rain because they think that the lightning strike is a stronger cause of the fire than the lack of rain. As mentioned earlier, people could think that the lightning strike is a stronger cause of the fire for a number of different reasons, including that the lightning strike is statistically abnormal, temporally recent, and connected through a physical process to the fire (Henne, Kulesza, et al., 2021; Icard et al., 2017; Kahneman & Miller, 1986; Wolff, 2007). But causes can also differ in the extent to which they contribute or could have contributed to an outcome, as in voting cases where some votes are weighed more heavily than others (Bernstein, 2017; Kaiserman, 2016; Kaiserman, 2018; Quillien & Barlev, 2021). Although we focus on normality in this paper, we call this explanation *the causal strength explanation* because it assumes that people have a concept of causation that distinguishes between strong and weak causes, regardless of the mechanism underlying this concept (e.g., dependence or production concepts) and regardless of which factors influence the perceived strength of a cause (e.g., normality, recency, contribution).

Importantly, however, there may be factors other than causal concepts that could induce gradation in causal judgments. Experiments on general causal judgments point to at least one important but underexplored possibility: participants' causal judgments are confounded by their degree of belief in that judgment (Collins & Shanks, 2006; Liljeholm & Cheng, 2009; Perales & Shanks, 2003; Schlottmann & Anderson, 1993; Shanks, 1987). For instance, when presented with several cases of individuals having headaches (or not) before and after consuming a dose of a mineral and asked, "To what degree does this mineral cause headaches?", both participants' causal judgments and their confidence in those judgments increased with the number of observations (Liljeholm & Cheng, 2009). The authors concluded that participants use causal judgments to indicate not only how strong they believe a causal relation is but also how confident they are that a causal relation exists in the first place. Independently, Sartorio (2020) used a similar line of reasoning to argue that degrees of causation are nothing but degrees of belief, concluding that causation itself may be non-graded. This explanation easily extends to *singular* causal judgments—judgments of whether a particular event caused a particular outcome—including the case of the lightning and the forest fire (Danks, 2017). Here, people may provide higher causal judgments of the lightning strike than of the lack of rain because they are more confident that the lightning strike caused the forest fire than they are that the lack of rain caused the fire. Under this explanation, which we refer to as the *confidence explanation*, people make graded causal judgments because they have varying degrees of belief in causal relations, regardless of whether they have graded or non-graded concepts of causation (Fig. 1, blue path).[1]

We stress that the confidence explanation is not mutually exclusive with the causal strength explanation; it could be the case that people have graded concepts of causation and graded degrees of belief in causal relations, both of which contribute to gradation in causal judgments. In line

[1] As pointed out by a reviewer, there is an important class of theories for which the line between causal strength and confidence in causal judgment is difficult to distinguish. Some dependence accounts of causal judgment assume that people's causal judgments are a function of their subjective degree of belief that the outcome would have been different if the cause had been different (e.g., Cheng, 1997; Gerstenberg et al., 2014; Spellman, 1997). If all it is for one event to cause another is that it makes a difference in this particular way, it may appear that these models are instances of the confidence explanation. However, we interpret them as providing measures of causal strength in terms of confidence in *counterfactual* relations, not necessarily *causal* relations, which classifies them as instances of the causal strength explanation. There are at least three reasons to treat confidence in counterfactual and causal relations as conceptually distinct. First, there is evidence that counterfactual cognition and causal cognition sometimes come apart: it has been shown, for instance, that counterfactual explanations require more cognitive effort than causal explanations (Byrne, 2016; McEleney & Byrne, 2006). Imagined counterfactuals also tend to focus on the preventability of controllable events, whereas causal ascriptions tend to focus on events that are thought to covary with the outcome (Mandel, 2003; Mandel & Lehman, 1996; N'gbala & Branscombe, 1995). If causal and counterfactual judgments come apart, then we have reason to treat them as conceptually distinct. Second, although some dependence theories do define causation in terms of a single class of counterfactual outcomes, most extant theories would argue either that more than one sort of counterfactual is relevant to causal judgments (e.g., Gerstenberg et al., 2021; Icard et al., 2017), that more than just counterfactuals are relevant to causal judgment (e.g., Hall, 2004; Lombrozo, 2010), or that counterfactuals in general are not relevant to causal judgment (e.g. Wolff, 2007). As we would like our discussion of gradation in causal judgments to apply to many theories of causal judgment, we do not want to make the assumption of equating confidence in a counterfactual relation and confidence in a causal relation. Finally, it is arguably normative for the strength and certainty of a causal relation to be kept distinct. Just as recent work in statistics has pushed for a division between indices of effect existence (e.g., the *p*-value) and indices of effect size (e.g., Cohen's *d*) under the observation of small-but-certain and large-but-uncertain effects (Cumming, 2014; Fritz, Morris, & Richler, 2012; Kruschke & Liddell, 2018; Makowski et al., 2019), we think that most theorists in causal judgment would not consider a causal relation to be stronger just because one is more certain in it. Nevertheless, some proponents of dependence theories may consider their accounts as instances of the confidence explanation if they are willing to equate confidence in a counterfactual relation with confidence in a causal relation.

with causal pluralism, it may also be that people make use of multiple (graded or non-graded) notions of causation which are combined to produce a graded judgment (Gerstenberg et al., 2021; Hall, 2004; Lombrozo, 2010). If changes in confidence are shown to account for *all* of the gradation in causal judgments, however, psychologists and philosophers may have good reason to favor extant accounts consistent with non-graded concepts of causation (Bello et al., 2018; Bello & Khemlani, 2015; Khemlani et al., 2014; Khemlani et al., 2018) and causation itself (Bernstein, 2017; Kaiserman, 2016; Kaiserman, 2018). Considering how many theories of causal judgment and causation hinge on whether people's concept of causation is graded, then, it is critical to (a) descriptively evaluate the extent to which people actually make graded singular causal judgments in the first place and to (b) evaluate the relative merits of the causal strength and confidence explanations of gradation in such judgments.

1.3. The present studies

Here, we sought to address this empirical gap by investigating the nature of gradation in singular causal judgments. To determine whether causal judgments are graded, we first qualitatively reanalyzed data from four previous experiments (Preliminary Analyses). When responding on a continuous scale, we found that although participants often provided graded judgments, such judgments tended to be less frequent than more categorical judgments at the ends of the scale. Given that causal judgments were sometimes graded, we tested two possible explanations for why this may be the case in two complementary experiments. Experiment 1 tested whether confidence in a causal judgment explains gradation in that judgment (the *confidence explanation*). In line with this explanation, we found that participants tended to give graded causal judgments when they were less confident, but tended to give more extreme causal judgments when they were more confident. In Experiment 2, we tested whether previously observed changes in normality, causal structure, and the number of candidate causes predicted causal judgments independently of confidence (the *causal strength explanation*). We found that although confidence explained some of the variation in causal judgments, it did not account for these classic effects on causal judgment. In sum, we found evidence that people do make graded causal judgments and that they make these judgments both because of gradation in their certainty in their judgments (in line with the confidence explanation) and because of gradation in their concept of causation (in line with the causal strength explanation).

2. Preliminary analyses

To assess the degree to which participants' judgments were graded, we first reanalyzed open data from four recent studies on singular causal judgments. In these studies, participants were first shown a stimulus (e.g., a written vignette or a video) where a candidate cause C occurs and then another event E occurs. Next, participants were asked the extent to which C caused E, either directly or indirectly through related notions, including responsibility or agreement with causal statements. Finally, participants made a judgment along some response scale (e.g., a Likert scale or slider scale). If causal judgments are not graded, one would expect that the majority of judgments would lie at the minimum and maximum of the scale, with only noise associated with measurement error in between. If causal judgments are graded, in contrast, one would expect that causal judgments frequently lie between the two extremes. We found that singular causal judgments were distributed multimodally at the middle and extremes of the scale, with fewer responses lying between these peaks.

2.1. Methods

We used open data from four published studies for this analysis chosen for the public availability and the size of their datasets (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Icard et al., 2017; Morris et al., 2019). Although we used a convenience sample, we note that these studies represent a broad range of conditions under which singular causal judgments are typically measured and under which gradation in these judgments is likely to occur. Specifically, they varied along types of stimuli (e.g., vignette- or video-based stimuli), types of response scales (e.g., 7-point Likert scales, 9-point Likert scales, or continuous slider scales), types of causes (e.g., generative, preventative, or omissive causes), and measures of causal strength (e.g., agreement with causal statements or causal responsibility). To determine how causal judgments were distributed, we compared histograms of causal judgments with data simulated from multilevel linear regressions assuming normally-distributed residuals, similar to the regression models used in the original studies. To accommodate the fact that causal judgments were made on a bounded scale, we truncated the simulated data to the range of the original data, fixing all values beyond this range to either the minimum or maximum value. All analysis code can be found online through the Open Science Framework (https://osf.io/dwjpt/).

2.2. Results

We depict histograms of causal judgments from the four studies, along with data simulated from multilevel linear regressions similar to those used in the original studies in Fig. 2. The same data are presented separately by experimental condition in Figs. S0.1-S0.5. There are two important patterns. First, participants' causal judgments were, in fact, graded. Participants not only responded at the scale extremes but also at most values between those extremes. Indeed, in Icard et al. (2017) and Morris et al. (2019), causal judgments at the center of the scale were about as frequent as responses at the scale extremes. In some cases (e.g., Fig. S0.1), the modal response was at the center of the scale. We take this to mean that causal judgments between the scale extremes do not strictly consist of noise; not only do participants intend to provide graded judgments, but at least in some cases, there is between-participant agreement on a particular graded response.

None of the four distributions, however, were particularly well-described by linear models assuming normally-distributed residuals. In each study as a whole (Fig. 2), and in many individual conditions within each study (Figs. S0.1-S0.5), distributions of causal judgments were multimodal. In other words, there were at least two peaks to each of the distributions, since responses at either end of the scale were more common than responses between each mode. While the regression models captured overall patterns in causal judgments across experimental conditions, they often predicted a higher prevalence of graded judgments than were actually present in the data. We note, however, that the regression models seem to better capture causal judgments made on response scales with many unique values (Gerstenberg et al., 2017; Morris et al., 2019) than on those with fewer unique values (Henne, Niemi, et al., 2019; Icard et al., 2017). We encourage readers to browse the condition-level histograms (Figs. S0.1-S0.5) for further information.

2.3. Discussion

In this small sample of studies, we found that distributions of singular causal judgments were graded but also multimodal. An immediate consequence of this result is that statistical and cognitive models that focus only on mean causal judgment (as opposed to distributions of causal judgments) may be misleading, because the mean does not always provide an informative metric of central tendency over multimodal distributions. In cases where the mean lies between two modes (e.g., the judgments for disjunctive causal structures in Fig. S0.2), for instance, standard linear models predict that judgments are most likely to lie at the mean. However, in these cases, this is actually the region where judgments are *least* likely to be made; definitionally, most judgments are located at either mode at the ends of the scale. This is potentially

problematic because the mean alone fails to differentiate between cases of between-participant *agreement* on a graded judgment and cases of between-participant *disagreement* with either widely-dispersed or largely non-graded judgments.

To the extent that psychological theories aim to account for gradation in causal judgments, then, they should also aim to account for this particular *kind* of gradation (i.e., gradation following a multimodal distribution). Existing theories fail in this regard, but they can in principle be extended given different assumptions about how causal judgments are generated (Cheng, 1997; Gerstenberg et al., 2021; Halpern & Hitchcock, 2015; Icard et al., 2017; Quillien, 2020). In line with recent work in other areas of psychology (e.g., Haines et al., 2020; Kennedy, Simpson, & Gelman, 2019; Schad, Betancourt, & Vasishth, 2021; Yarkoni & Westfall, 2017), we believe that such extensions will, in addition to fitting the data better, help to provide a deeper understanding of how causal judgments are made. As such, we think that the challenge of modeling individual causal judgments is a promising avenue of future research for models that are already successful at the group level (Gerstenberg et al., 2021; Icard et al., 2017; Quillien, 2020). In the meantime, to deal with this potential difficulty, for each of our experiments below we display raw distributions of causal judgments, report group-level (i.e., random) effects identifying differences in trends among participants and stimuli, and compare several statistical models of our data with varying degrees of complexity under a Bayesian framework (Vehtari, Gelman, & Gabry, 2017).

3. Experiment 1

In our Preliminary Analyses, we found that although causal judgments were multimodal, many judgments were graded. In our first experiment, we sought to determine the utility of the confidence explanation by investigating whether confidence explained any gradation in causal judgments. We hypothesized, in line with the confidence explanation, that participants would make intermediary causal judgments to express uncertainty in an event's causal role. In particular, the confidence explanation predicts an interaction between causal strength and confidence: when participants are uncertain, they will make graded causal judgments, and when they are certain, they will make non-graded causal judgments at the minimum (e.g., when they think that X did not cause Y) or maximum (e.g., when they think that X caused y) of the scale.

We presented participants nine vignettes (including the example shown in Table 1) taken from Icard et al. (2017) and Henne, Niemi, et al. (2019) twice in a matched-vignette design. During one presentation of each vignette, we recorded causal judgments using a categorical scale (did not cause, partially caused, or totally caused) as a discrete measure of perceived causal strength. We then measured confidence in this categorical causal judgment to indicate participants' degree of certainty. On a repeated presentation of the same vignettes with only the names of characters changed, we also recorded causal judgments from the same participants on a typical continuous scale. Finally, we used the discrete causal judgments and the confidence in those discrete judgments from one presentation of the vignette to predict continuous causal judgments on the repeated presentation. If the confidence explanation is false, we should see no particular relation between participants' confidence in their discrete causal judgment and their continuous causal judgment. Rather, we should see that the only predictor of gradation in participants' continuous causal judgments is their discrete causal judgment. On

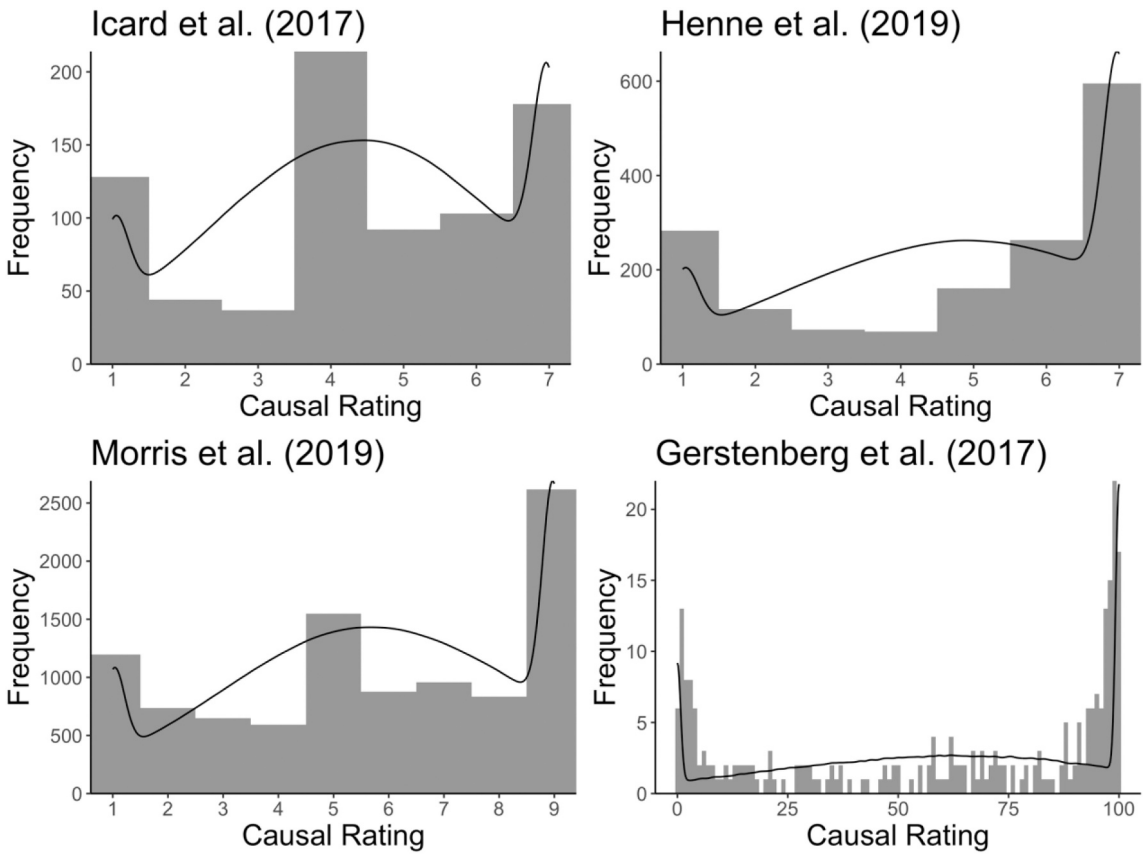

Fig. 2. Causal judgments are graded but multimodal.
Histograms of singular causal judgments from the four representative studies used in the Preliminary Analyses. Solid lines indicate simulated data from multilevel linear regressions similar to those used in the original studies (see SI for further details for each study). While distributions of causal judgments appear to be graded, models assuming normally distributed residuals do not account for the multimodality present in the data.

Table 1
Example vignette used in Experiment 1.

1a) Conjunctive: Billy and Suzy inherited an unusual type of hybrid car that has two special car batteries called Bartlett batteries. The car won't start unless it has two Bartlett batteries. Having one battery isn't enough to start the car. When they got the car, both Bartlett batteries were missing.	*1b) Disjunctive:* Billy and Suzy inherited an unusual type of hybrid car that has two special car batteries called Bartlett batteries. The car won't start unless it has at least one Bartlett battery. Having a second Bartlett battery isn't necessary to start the car. When they got the car, both Bartlett batteries were missing.
2a) No violation: One day, Billy and Suzy are both out of the house. Billy is visiting his friend's house, and notices that his friend has a Bartlett battery. Billy asks his friend to sell the battery to him, and his friend says that he's willing to sell it for a fair price, so Billy buys the Bartlett battery from his friend.	*2b) Norm violation:* One day, Billy and Suzy are both out of the house. Billy is visiting his friend's house, and notices that his friend has a Bartlett battery. Billy asks his friend to sell the battery to him, but his friend says that he can't sell it because he needs it for his own car. Billy waits until his friend is in the bathroom, and then steals the Bartlett battery from his friend.
Meanwhile, on the other side of town, Suzy walks into an automotive parts shop and happens to notice that they have a single Bartlett battery in stock. Suzy decides to buy the Bartlett battery from the shop. When Billy and Suzy get home, they installed the two Bartlett batteries.	
1a) Conjunctive (con't): Since the car now had both Bartlett batteries, they were able to start the car.	*1b) Disjunctive (con't):* Since all the car needed was at least one Bartlett battery, they were able to start the car.

Battery vignette used in Experiment 1, taken from Icard et al. (2017). Vignette can use either a conjunctive (*1a*) or disjunctive (*1b*) causal structure and can either involve a norm violation (*2b*) or no norm violation (*2a*). In the repeated presentation of the vignette, the names "Billy" and "Suzy" were replaced with "Alex" and "Laurie".

the other hand, if the confidence explanation is true, we should see an interaction between confidence and discrete causal judgment. In particular, we should find that the less confident participants are in their discrete causal judgments, the more likely they are to make a graded causal judgment when given a continuous scale. In this way, our matched-vignette design allowed us to directly evaluate the confidence explanation of gradation in causal judgments conditional on whether or not the participant thought that X caused Y. In line with the confidence explanation, we hypothesized that participants' continuous causal judgments would be predicted by an interaction between their discrete causal judgments and their degree of confidence in the discrete judgments.

3.1. Methods

3.1.1. Participants

80 participants were recruited from Amazon Mechanical Turk. Pilot data showed that this sample size would be large enough to detect our effects of interest. We report as meaningful any Bayesian credible interval that exceeds a standardized effect size of .1 (see Analysis). All participants were from the United States, had an overall HIT approval rate of at least 99%, received $4.50 in compensation after completing the task, and provided informed consent in accordance with Duke University IRB. Data from two participants were excluded because they reported not paying attention to the task (see Supplementary Information p. 6). Data were analyzed from the remaining 78 participants (M_{age} = 37, SD_{age} = 10.39, 33 female, 44 male, 1 other).

3.1.2. Materials

Stimuli were nine vignettes involving two independent candidate causes that combine in a conjunctive (i.e., both causes are necessary, neither is individually sufficient) or disjunctive (i.e., both causes are individually sufficient, neither is necessary) manner to produce an effect. An example vignette is shown in Table 1. Six of these vignettes were used in Icard et al. (2017), and they varied by whether the focal cause was statistically or prescriptively normal or abnormal. The other three vignettes were used in Experiment 2 of Henne, Niemi, et al. (2019), and they varied by whether the focal cause was an action or inaction. Thus, we manipulated two aspects of the vignettes between participants: causal structure (conjunctive vs. disjunctive) and either normality (normal vs. abnormal) or action (action vs. inaction). Importantly, in this experiment, we were not interested in changes in causal judgments due to these manipulations because we had no hypotheses about the relationship between confidence, normality, and causal structure (investigating which is the aim of Experiment 2). As these manipulations serve merely to generate variability in causal judgments, we collapsed across them in our analyses. However, our results are depicted broken down by vignette and by experimental condition in Figs. S1.1 and S1.2, respectively. All stimuli are available in the Supplementary Information, and all materials, data, and code are available via the Open Science Framework (https://osf.io/dwjpt/).

3.1.3. Procedure

In a 2 (Rating scale: Discrete vs. Continuous) x 9 (vignette) within-participants design, participants read nine vignettes twice in randomized order. The names of the characters in all vignettes were randomized to either "Billy" and "Suzy" in one presentation and "Alex" and "Laurie" in the other. After one presentation of each vignette, participants responded to the prompt "X __ Y" on a discrete scale (i.e., "did not cause", "partially caused", "totally caused"), where X is the focal cause and Y is the outcome of the vignette. After the other presentation of the same vignette, participants responded to the question "To what extent did X cause Y?" on a continuous scale ranging from "not at all" (coded as a numerical response of 0) to "totally" (coded as a numerical response of 1) with no labeled midpoints. Presentation order of each rating scale was randomized, so that for some vignettes the discrete scale was presented first and for others the continuous scale was presented first. After each causal judgment, participants were asked "How confident is your response?" and provided a confidence rating on a continuous scale ranging from "not at all" confident to "totally" confident with no labeled midpoints. In this paper we focus on confidence in participants' discrete causal judgments, though we found similar results with confidence in continuous causal judgments (see section *Confidence ratings are reliable across time and measurement scales*). To reduce the possibility that participants simply attempted to reproduce their responses on the second presentation, all vignettes were presented once before any vignettes were presented for a second time, presentation order was randomized, and the names of the characters in each vignette were randomized. Data from all continuous scales were rescaled to the [0, 1] range (where 0 indicates not at all causal/confident and 1 indicates totally causal/confident).

3.1.4. Analysis

To determine the effect of confidence on causal judgments, we used the *brms* package in R to fit a multilevel Bayesian linear regression model over participants' continuous causal judgments (Bürkner, 2017). Multilevel Bayesian regression models are similar to frequentist regressions, except that they provide posterior distributions of model coefficients and model predictions. As a result, Bayesian regression models provide more interpretable estimates of uncertainty (Kruschke, 2011). We utilized multilevel Bayesian regressions to predict participants' continuous causal judgments using their discrete causal judgments (i.e., "did not cause", "partially caused", "totally caused") and their confidence in the discrete causal judgments (in the [0-1] range) from the matched vignette

as predictors. We included correlated random intercepts for each participant and for each vignette. We used four MCMC chains with 2500 iterations of sampling and 2500 iterations of burn-in, weakly-informative Gaussian priors with a mean of 0.0 and a standard deviation of 1.0 for all coefficients on the mean causal judgment, as well as weakly-informative Gaussian priors with a mean of 0.0 and a standard deviation of 5.0 for all coefficients on the standard deviation of causal judgments. In order to ensure that estimated SDs remain positive, they were estimated with a log link, and differences in SD were tested on the linear scale of the model coefficients.

Following the conventions of Bayesian parameter estimation (Kruschke & Liddell, 2018), we report posterior medians, 95% Highest Density Intervals (*HDI*s), a Bayesian analog of the frequentist *p*-value (*p-pd*), and percentage of the 95% *HDI*s inside a Region of Practical Equivalence (ROPE) corresponding to a standardized effect size of [-.1, .1] for each coefficient in our model. The Bayesian *p*-value we report is a linear transformation of the probability of direction (which is defined as the proportion of posterior values greater than or less than 0) that shares a similar interpretation to the frequentist *p*-value in that a value less than .05 indicates the existence of an effect (Makowski, Ben-Shachar, Chen, & Lüdecke, 2019). The percentage of the 95% *HDI* inside a ROPE indicates the significance of an effect's size and is defined as the proportion of the 95% most probable posterior values inside a null region, where $< 5\%$ is a rough benchmark for rejecting the null (Kruschke, 2011).

A significant advantage of the Bayesian approach is that it allows us to model a wide variety of distributions, and it provides robust methods to adjudicate between such models (Vehtari et al., 2017). Here we tested four models. The simplest model was a standard linear regression assuming normality, linearity, and homogeneity of variance, as is common in the literature on singular causal judgments (e.g., Gerstenberg et al., 2017; Henne, Bello, et al., 2019; Icard et al., 2017; Morris et al., 2019). Next, we augmented this model to allow for heterogeneity of variance, modeling both the mean and the standard deviation of causal judgments for each experimental condition. We also tested a zero-one-inflated Beta (ZOIB) regression model, which models the probability that a judgment is graded, models graded judgments using a Beta distribution constrained to (0, 1), and models non-graded judgments exactly at 0 or 1 using a Bernoulli distribution. Finally, we also tested a generalized additive model which accounts for possible non-linearities in the relationship between confidence and causal judgments. Each of these models included main effects and an interaction between discrete causal judgment and confidence in that judgment, as well as correlated random intercepts, for each distributional parameter in the model.

3.2. Results

3.2.1. Causal judgments are multimodal and heteroscedastic

We focus below on confidence in participants' discrete causal judgments, though results for confidence in participants' continuous causal judgments were similar (see section *Confidence ratings are reliable across time and measurement scales*). As found in the Preliminary Analyses, the distribution of continuous causal judgments was multimodal, with peaks at the bottom, center, and top of the scale (Fig. 3*A*). As predicted, each of these modes (at 0, 0.5, and 1.0) corresponded to a particular discrete causal judgment (did not cause, partially caused, totally caused, respectively), reflecting that most variation in the continuous causal judgments was captured by the discrete judgments. To best describe these judgments, we compared four different models (see Methods). As compared by PSIS-LOOIC and model stacking weights, the unequal variance Gaussian model provided the best tradeoff between model complexity and fit (Table 2*A*). All further analyses are performed with this model, for which posterior predictions and group-level effects are presented in Figs. 3B–D and S1.3, respectively. These results indicate that although our data would not be well-described by a standard regression model, allowing for nonlinear relationships between causal judgments and confidence did not significantly improve model fit.

3.2.2. Low-confidence causal judgments are graded

If the confidence explanation (i.e., that gradation in causal judgments is due to uncertainty) is correct, we should find that graded causal judgments are made with low confidence, whereas judgments at the scale extremes are made with high confidence, manifesting as a significant interaction between the discrete causal judgment and the confidence in that judgment when predicting participants' continuous causal judgments. If, however, the confidence explanation is incorrect, then causal judgments should be unrelated to confidence. We found support for the confidence explanation: there was strong evidence for an interaction between discrete causal judgment and confidence (Table 3C, Fig. S1.6), and model comparisons revealed that this interaction was needed to best account for the data (Table S1.1A). While judgments of candidate causes deemed partially causal did not change with confidence, judgments of candidate causes deemed totally causal increased with confidence, and judgments of candidate causes deemed non-causal decreased with confidence (Table 3A, Fig. S1.4). Pairwise contrasts confirmed that while causal judgments between candidate causes categorized as non-causal, partially causal, and totally causal were indistinguishable at low confidence, they were separable at high confidence (Table 3B, Fig. S1.8). Posterior estimates and coefficients of mean causal judgments are plotted in Fig. 3C and Fig. S1.10, respectively.

3.2.3. Low-confidence causal judgments are more variable

If the confidence explanation is correct, we might also expect to find that judgments with low confidence are more variable, since these uncertain judgments are supposedly less reliable. As our model allows the standard deviation of continuous causal judgments to vary between conditions, we can also test for these differences. There was strong evidence that continuous causal judgments became less variable with increasing confidence for causes deemed partially causal or totally causal, but not for those deemed non-causal (Table 3D, Fig. S1.5). Additionally, this decrease in variability in causal judgments with increasing confidence was stronger for causes deemed totally causal than those deemed partially causal, but there was no evidence for a difference in the effect of confidence between candidate causes deemed non-causal and partially causal (Table 3F, Fig. S1.7). Pairwise contrasts revealed that at low confidence (confidence $= 0$), there was no evidence for a difference in variability between judgments of candidate causes deemed totally causal, partially causal, and non-causal. In contrast, at high confidence (confidence $= 1$), there was strong evidence that judgments for causes deemed totally causal were less variable than those rated as partially causal, and that causes rated as non-causal were more variable than those rated as partially causal (Table 3E, Fig. S1.9). Posterior estimates and coefficients of the standard deviation of causal judgments are plotted in Fig. 3D and Fig. S1.11, respectively.

3.2.4. Confidence ratings are reliable across time and measurement scales

As exploratory analyses, we wanted to see whether participants' confidence in their discrete causal judgment resembled their confidence in their continuous causal judgments on the matched presentation of each vignette. Descriptively, the majority of both continuous and discrete causal judgments were highly confident: only 44 (6.3%) of discrete causal judgments and 54 (7.7%) of continuous causal judgments were made with confidence less than 50%. Furthermore, we found that these two confidence ratings were moderately to highly correlated ($r = .46$, 95% *HDI* $= [.40, .52]$, *p-pd* $< .001$, *ROPE* $= [-.1, .1]$, 0% in *ROPE*). To strengthen this claim, we decided to replicate our above results using participants' confidence in the continuous causal judgment in place of their confidence in the discrete causal judgment as a predictor, and we found the same qualitative pattern of results, which can be found in the Supplementary Information (Table S1.1B, Fig. S1.12).

3.3. Discussion

The results from Experiment 1 demonstrated that, in line with the

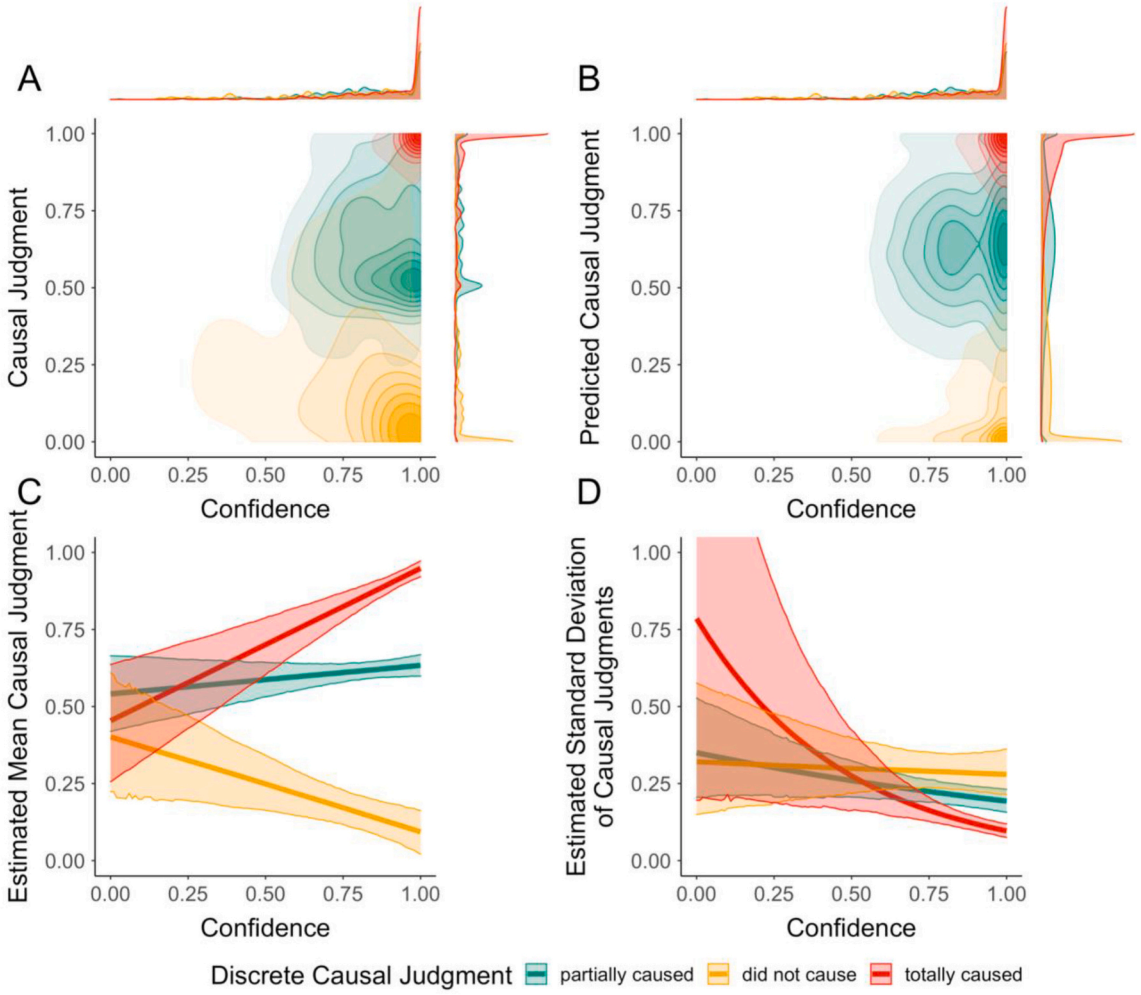

Fig. 3. Causal judgments are modulated by confidence.
(*A*) Density plots of distributions of causal judgments versus confidence. Participants' continuous causal judgments aligned with discrete causal judgments but were also modulated by confidence. (*B*) Same as (*A*) for predictions from the fitted model. The model recapitulated qualitative features of the distribution. (*C*) Causal judgments increased with confidence for causes deemed totally causal, decreased with confidence for causes deemed non-causal, and were unrelated to confidence for causes deemed partially causal. (*D*) Variability in continuous causal judgments generally decreased with confidence. Lines indicate posterior medians, and shaded areas reflect 95% HDIs. Posterior predictions are restricted to the [0, 1] range for visualization, but may in fact exceed these bounds.

Table 2
Causal judgments were best explained by unequal variance normal distributions.

Model	ELPD	PSIS-LOOIC	Effective # of parameters	ΔELPD	Model weight
A) Experiment 1					
UV Gaussian	105.8 (*30.8*)	-211.6 (*61.6*)	128.9 (*10.8*)	0.0 (*0.0*)	.797
EV Gaussian	34.3 (*25.7*)	-68.6 (*51.4*)	49.4 (*3.6*)	-71.5 (*17.8*)	.192
GAM	25.8 (*26.3*)	-51.6 (*52.2*)	60.5 (*5.1*)	-80.0 (*18.1*)	.000
ZOIB	-149.0 (*28.5*)	298.0 (*57.0*)	172.2 (*9.2*)	-254.8 (*26.8*)	.011
B) Experiment 2					
UV Gaussian	-78.1 (*67.6*)	156.2 (*135.2*)	123.9 (*7.7*)	0.0 (*0.0*)	.72
EV Gaussian	-116.2 (*61.7*)	232.4 (*123.3*)	45.6 (*1.6*)	-38.1 (*20.0*)	.28
ZOIB	-2432.5 (*55.9*)	4865.0 (*111.8*)	179.7 (*10.6*)	-2354.4 (*43.5*)	.0

Posterior means and standard errors of model comparisons between equal variance (EV) Gaussian, unequal variance (UV) Gaussian, generalized additive (GAM), and zero-one inflated Beta (ZOIB) models of the relationship between confidence and causal judgments for Experiment 1 (*A*) and Experiment 2 (*B*). Models are listed in descending order of performance. In both experiments, causal judgments were best predicted by the unequal variance Gaussian model. Model weights were computed using model stacking. ELPD = expected log predictive density, PSIS-LOOIC = pareto-smoothed importance sampling leave one out information criteria.

confidence explanation, participants' continuous causal judgments could be predicted as a function of a discrete causal judgment and the degree of confidence in that judgment. In this experiment, participants gave a discrete causal judgment (i.e., 'did not cause', 'partially caused', 'totally caused') and rated their confidence in that causal judgment in response to a vignette. They also gave continuous causal judgments in response to matched vignettes with only the names of characters changed. When participants were uncertain about their discrete causal judgments, their corresponding continuous causal judgments were graded and highly variable. Regardless of whether the participant

Table 3
Contrasts of the mean and standard deviation of causal judgments from Experiment 1.

	Parameter	Discrete causal judgment	Confidence	Median	95% HDI	*p-pd*	ROPE	% in ROPE
A	Mean	PC	High–Low	0.09	[−0.06, 0.24]	.204	[−0.03, 0.03]	16.89
	Mean	DNC	High–Low	−0.31	[−0.54, −0.09]	.006	[−0.03, 0.03]	0
	Mean	TC	High–Low	0.5	[0.31, 0.71]	< .001	[−0.03, 0.03]	0
B	Mean	DNC - PC	Low	−0.14	[−0.36, 0.09]	.24	[−0.03, 0.03]	11.18
	Mean	TC - PC	Low	−0.09	[−0.32, 0.13]	.43	[−0.03, 0.03]	16.74
	Mean	DNC - PC	High	−0.54	[−0.61, −0.46]	< .001	[−0.03, 0.03]	0
	Mean	TC - PC	High	0.32	[0.27, 0.36]	< .001	[−0.03, 0.03]	0
C	Mean	DNC - PC	High–Low	−0.4	[−0.67, −0.14]	.003	[−0.03, 0.03]	0
	Mean	TC - PC	High–Low	0.41	[0.17, 0.66]	< .001	[−0.03, 0.03]	0
D	SD	PC	High–Low	−0.59	[−1.13, −0.06]	.026	[−0.06, 0.06]	0.18
	SD	DNC	High–Low	−0.13	[−0.92, 0.61]	.73	[−0.06, 0.06]	12.77
	SD	TC	High–Low	−2.11	[−3.15, −1.02]	< .001	[−0.06, 0.06]	0
E	SD	DNC - PC	Low	−0.08	[−0.81, 0.63]	.81	[−0.06, 0.06]	14.10
	SD	TC - PC	Low	0.82	[−0.13, 1.91]	.12	[−0.06, 0.06]	3.13
	SD	DNC - PC	High	0.37	[0.09, 0.63]	.008	[−0.06, 0.06]	0
	SD	TC - PC	High	−0.69	[-0.96, −0.43]	< .001	[−0.06, 0.06]	0
F	SD	DNC - PC	High–Low	0.45	[−0.46, 1.33]	.33	[−0.06, 0.06]	6.72
	SD	TC - PC	High–Low	−1.52	[−2.71, −0.35]	.012	[−0.06, 0.06]	0

(*A,D*) The effect of confidence on continuous causal judgments for the three levels of discrete causal judgment. (*B,E*) Contrasts of continuous causal judgments between discrete causal judgments for low confidence and high confidence. (*C,F*) Interaction contrasts testing for differences in the effect of confidence on continuous causal judgments between the three discrete causal judgments. Whereas low-confidence causal judgments were graded, high-variance, and unrelated to discrete causal judgments, high-confidence judgments were multi-modal, low-variance, and clustered according to discrete causal judgments. SD = standard deviation, DNC = "did not cause", PC = "partially caused", TC = "totally caused", HDI = highest density interval, *p-pd* = probability-of-direction-based *p* value, ROPE = region of practical equivalence.

ranked the candidate cause as non-causal, partially causal, or totally causal, if participants were uncertain about their ranking, their corresponding continuous causal judgments were intermediary and indistinguishable. In contrast, when participants were highly confident about their discrete causal judgments, their corresponding continuous causal judgments were multimodal and less variable. These continuous causal judgments were high when participants confidently ranked the candidate cause as totally causal, low when participants confidently ranked them as non-causal, and intermediate when participants confidently ranked them as partially causal.

These results have both methodological and theoretical consequences. In terms of methodology, our findings verify that continuous causal judgments are non-normally distributed and that distributions of causal judgments tend to vary not only in location (i.e., mean), but also in dispersion (i.e., variance). Although response distributions of this form have appeared in data from several studies on causal judgment to-date (see Preliminary Analyses), currently most work ignores differences in variability and even assumes equal variances in accordance with standard linear models. While this practice may be sufficient to make group-level inferences about causal judgments, we found that modeling variance in causal judgments provided better descriptions of the data. Following this finding, we recommend three practical suggestions for other researchers working in causal cognition: (a) plot histograms of causal judgments, (b) compare multiple statistical models of judgments, and (c) model distributions (and not simply the mean) of judgments using a generative approach (Haines et al., 2020; Kennedy et al., 2019; Schad et al., 2021). Although more work is needed to identify response distributions that are capable of reproducing causal judgments, in the meantime, these practices will help ensure that researchers interpret their results in a manner that is consistent with their data.

In terms of theory, these results provide strong support for the confidence explanation of gradation in causal judgments. We found that continuous causal judgments reflected an interaction of discrete causal judgments and confidence. That is, judgments with low confidence tended to be more variable and more graded, whereas judgments with high confidence tended to be less variable and less graded. This finding that confidence helps to explain gradation in causal judgment potentially presents a challenge to existing research operating under the causal strength explanation. Notably, if people sometimes give graded judgments of candidate causes to indicate that they are uncertain about the event's causal role, it is possible that they *only* make graded judgments for this reason. In other words, differences in causal judgments that appear to be due to causal strength may in fact arise from differences in confidence, and may disappear when controlling for confidence. If confidence can explain away other effects on causal judgment, such a result would indicate that people have a non-graded concept of causation, but nevertheless make graded causal judgments according to their degree of belief in the causal relation. In Experiment 2, we sought to answer this question in a more controlled setting.

4. Experiment 2

As we saw with the case of the lightning and the forest fire, causal judgments are sensitive to normality, among other factors (Henne et al., 2017; Henne, O'Neill, et al., 2021; Icard et al., 2017; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Morris et al., 2019). The causal strength explanation interprets these results as meaning that people adjust their beliefs about how causal an event is based on how normal it is: abnormal events are seen as more causal than normal ones. The results from Experiment 1, in line with the confidence explanation, indicate that uncertainty is also associated with gradation in causal judgments. This new finding suggests two possibilities. If the variation in causal judgments due to normality is unexplained by confidence, then we can say that both the causal strength explanation and the confidence explanation hold. That is, people have a graded concept of causation, and they make graded causal judgments not only to indicate the strength of a cause, but also to indicate their level of certainty in that causal judgment. On the other hand, if confidence fully mediates the effect of normality on causal judgments, we may have reason to doubt the causal strength explanation. In this case, people might have a non-graded concept of causation and nevertheless rate abnormal causes as higher than normal causes just because they are more confident that they are causal, not because they are represented as being more causal. In Experiment 2, we directly test the causal strength explanation by seeing whether these well-documented effects on causal judgments are mediated by confidence.

To investigate this question, we focused on two robust normality effects on causal judgment illustrated by the example vignette in Table 4:

Suppose that Doug, John, Jack, and Kris each own local businesses sharing a sewer system that can process waste from one business on a given day; waste from two businesses is enough to contaminate the system. One day, Doug and John both dump waste from their businesses. Consequently, the system gets contaminated with waste.

In this conjunctive structure (i.e., both businesses had to dump waste to contaminate the system), people often exhibit *abnormal inflation*: that is, they typically judge that Doug dumping waste was more of a cause of the contamination when he violated a norm in doing so (e.g., by dumping the waste when he is not allowed) than when he did not violate a norm (e.g., by dumping the waste on an allowed day). Now assume that the waste from just one business is enough to contaminate the system. In this disjunctive structure, people instead exhibit *abnormal deflation*: Doug dumping his waste is often judged as *less* of a cause of the contamination when he violated a norm than when he did not violate a norm (Icard et al., 2017).

In addition to causal structure and normality, we also manipulated the number of candidate causes in each vignette by allowing between one and four causes to produce an effect. When only one cause occurs (i.e., when Doug was the only one to dump waste and the system was contaminated thereafter), the conjunctive and disjunctive structures look the same, serving as a baseline for comparison. As such, we expected participants to rate Doug as totally and solely causal in these conditions. When four causes are present (i.e., when Doug, John, Jack, and Kris all dump waste on the same day), however, we expected people to make more graded causal judgments of Doug, since he was only one of four contributing factors. In addition to providing a natural manipulation of causal strength, this allowed us to generalize abnormal inflation and deflation to cases with only one or more than two candidate causes. Finally, we tested effects on the mean and variance of both causal judgments and participants' confidence in them. As in Experiment 1, we interpret the variance in causal judgments as an indicator of the extent to which participants agree on the mean causal judgment: low variance suggests high agreement between participants, and high variance suggests low agreement. In contrast, we estimated changes in the variance of confidence ratings to better detect changes in these ratings. Specifically, in Experiment 1 we saw that participants almost exclusively exhibited high confidence in their causal judgments. Due to the presence of this ceiling effect, we determined that changes in the variance of confidence ratings might provide a more sensitive measure of overall increases or decreases in confidence than the mean alone.

To be precise, we designed Experiment 2 to test whether changes in causal judgments due to normality, causal structure, and the number of causes could be explained by changes in confidence in causal judgments. If confidence explains these effects, then we have reason to abandon the causal strength explanation: people have non-graded concepts of causation but they nevertheless make graded causal judgments when they are uncertain. If confidence does not explain these effects, we can accept both the confidence explanation and the causal strength explanation: people have graded concepts of causation and they make graded causal judgments both when they are uncertain and when they believe a cause is weak.

4.1. Methods

4.1.1. Participants

To roughly match the statistical power of the two experiments that established the abnormal inflation and abnormal deflation effects, we based recruitment on a target of 1920 participants, or 20 participants for each condition and each vignette (Icard et al., 2017). All participants were recruited from Amazon Mechanical Turk, from the US, had an overall HIT approval rate of at least 95%, received $0.50 in compensation after completing the task, and provided informed consent in accordance with Duke University IRB. Data from 116 (5.7%) additional participants were collected but excluded and replaced because they reported not paying attention to the task (Rouder, 2014; see Supplementary Information p. 6). Data were analyzed from the remaining 1920 participants ($M_{age} = 37$, $SD_{age} = 11.63$, 834 female, 1081 male, 4 other).

4.1.2. Materials

Stimuli were six vignettes involving one, two, three, or four candidate causes in conjunctive or disjunctive causal structures (Table 4). We varied the normality of the focal cause (the candidate cause that participants were asked to rate) using either a descriptive (3 vignettes) or prescriptive (3 vignettes) norm. We included both descriptive and prescriptive norms to ensure that our results would generalize to both cases, since the normality effects of interest have been shown to apply to both

Table 4
Example vignette used in Experiment 2.

Doug, John, Jack, and Kris own local businesses that share a sewer system. Recently, there has been a lot of rain, so the water treatment plant is stressed.				
C 1: Given their current demands, the water treatment plant cannot decontaminate waste from any businesses. If one local business dumps their waste, the water processed by the treatment plant will be contaminated with waste.	*C 2:* Given their current demands, the water treatment plant can only decontaminate so much waste. If two local businesses dump their waste on the same day, the water processed by the treatment plant will be contaminated with waste.	*C 3:* Given their current demands, the water treatment plant can only decontaminate so much waste. If three local businesses dump their waste on the same day, the water processed by the treatment plant will be contaminated with waste.	*C 4:* Given their current demands, the water treatment plant can only decontaminate so much waste. If four local businesses dump their waste on the same day, the water processed by the treatment plant will be contaminated with waste.	*D:* Given their current demands, the water treatment plant cannot decontaminate waste from any businesses. If any local businesses dump their waste, the water processed by the treatment plant will be contaminated with waste.
Normal: Doug, John, Jack, and Kris are all allowed to dump their waste water on Mondays.		*Abnormal:* Doug is not allowed to dump his businesses waste water on Mondays. John, Jack, and Kris are all allowed to dump their waste water on Mondays.		
1: This Monday Doug dumped his business's waste water.	*2:* This Monday Doug dumped his business's waste water. On the same day, John also dumped his waste water.	*3:* This Monday Doug dumped his business's waste water. On the same day, John also dumped his waste water, and Jack dumped his.	*4:* This Monday Doug dumped his business's waste water. On the same day, John also dumped his waste water, Jack dumped his, and Kris dumped his.	
C 1: Sure enough, since one person dumped their business's waste water, the water processed by the treatment plant was contaminated with waste.	*C 2:* Sure enough, since two people dumped their business's waste water, the water processed by the treatment plant was contaminated with waste.	*C 3:* Sure enough, since three people dumped their business's waste water, the water processed by the treatment plant was contaminated with waste.	*C 3:* Sure enough, since four people dumped their business's waste water, the water processed by the treatment plant was contaminated with waste.	*D:* Sure enough, since at least one person dumped their business's waste water, the water processed by the treatment plant was contaminated with waste.

Sewer vignette used in Experiment 2. This vignette can use either a conjunctive (C) or disjunctive (D) causal structure and can either involve a norm violation (Normal) or no norm violation (Abnormal) and anywhere from 1 to 4 contributing candidate causes.

types of norms (Icard et al., 2017; Kominsky & Phillips, 2019). However, we did not predict any differences between vignettes with different types or norms, so we collapsed across these conditions in our analyses, though we present results separated by norm type in Fig. S2.3. Thus, all stimuli had three components that were manipulated between subjects: causal structure (conjunctive vs. disjunctive), normality (normal vs. abnormal), and number of candidate causes (1–4). All stimuli are available in the Supplementary Information, and all materials, data, and code are available via the Open Science Framework (https://osf.io/dwjpt/).

4.1.3. Procedure

In a 2 (Causal Structure: Conjunctive, Disjunctive) x 2 (Normality: Normal, Abnormal) x 4 (Number of Candidate Causes: 1, 2, 3, 4) x 6 (vignette) between-participants design, participants each read a single vignette. In each vignette, participants were informed of four possible causes of some effect. Participants were then told that at least one, two, three, or four of these events need to occur for the effect to occur. Next, participants were told that a focal cause either violated a norm (Abnormal) or not (Normal) and that the remaining causes did not violate any norm. Finally, participants were told that either one, two, three, or all four of these candidate causes actually occurred, that the remaining candidate causes did not occur, and that the effect occurred. In the Disjunctive conditions, one cause was always sufficient to bring about the effect. In the Conjunctive conditions, the number of causes needed for the effect to occur was always the same as the number that actually occurred. After reading the vignette, participants responded to the question "To what extent did X cause Y?" on a continuous scale ranging from "not at all" (coded as 0) to "totally" (coded as 1), where X is the focal candidate cause. After each causal judgment, participants provided a confidence rating in response to the question "How confident are you in your response?" on a continuous scale ranging from "not at all" (coded as 0) confident to "totally" (coded as 1) confident. Data from all scales were rescaled to the [0, 1] range (where 0 indicates not at all causal/confident and 1 indicates totally causal/confident).

To ensure that all of our manipulations were between-participants, we decided not to collect discrete causal judgments in this experiment. This change may have limited the extent to which confidence could explain the effects of normality, causal structure, and number of candidate causes, since in Experiment 1 we found that the relationship between confidence and causal judgment was contingent on the participants' discrete causal judgment. However, we also found in Experiment 1 that participants were relatively unlikely (17.8%) to select "did not cause" because our vignettes did not ask about events that were clearly not at all causes of the outcome. Nonetheless, since confidence had a positive relationship with causal judgments both when participants selected "totally caused" or "partially caused," confidence had a positive relationship with causal judgments overall. So, we reasoned that confidence alone might still be able to account for effects on causal judgment without use of the discrete measure.

4.1.4. Analysis

To determine the effects of normality, causal structure, and number of candidate causes on causal judgments and confidence, we used the *brms* package in R to fit multivariate multilevel Bayesian linear regression models over participants' causal judgments and confidence ratings (Bürkner, 2017). Specifically, this model predicted the mean and standard deviation of causal judgments and confidence from normality, causal structure, and the number of candidate causes. The model included correlated random intercepts by vignette. Finally, the model also estimated the residual correlation between causal judgments and confidence after taking into account the variance explained by normality, causal structure, and the number of candidate causes, which was used to calculate the proportion of the effects mediated by confidence. Unless reported, all analyses were performed as in Experiment 1.

4.2. Results

4.2.1. Causal judgments are multimodal and heteroscedastic

As found in the Preliminary Analyses and in Experiment 1, the marginal distribution of continuous causal judgments was multimodal (Fig. 4*A*). To determine how to best model these judgments, we tested a standard linear regression assuming equal variances of causal judgments, an unequal variance Gaussian regression, and a zero-one-inflated beta regression. As compared with PSIS-LOOIC and model stacking weights, the unequal variance Gaussian model once again provided the best tradeoff between model complexity and fit (Table 2*B*). We performed all remaining contrasts on this model, for which posterior predictions and vignette-level effects are plotted in Figs. 4B and S2.3, respectively. Posterior estimates of the mean and standard deviation of causal judgments and confidence are presented in Figs. 5 and 6. These results provide further support that unequal variance Gaussian models provide a good balance between model fit and model complexity.

4.2.2. Abnormal inflation generalizes to many causes

First, we investigated the extent to which causal judgments in conjunctive structures were higher for abnormal causes than for normal candidate causes (i.e., *abnormal inflation*). Although there was no evidence for abnormal inflation in vignettes with one candidate cause, there was strong evidence for abnormal inflation in vignettes with two, three, and four candidate causes (Fig. S2.6). There was also strong evidence that causal judgments were less variable for abnormal causes in vignettes with any number of candidate causes (Fig. S2.7). Finally, we tested whether abnormal inflation occurred in confidence ratings of these causal judgments. There was no evidence for normality-driven differences in mean confidence for vignettes with any number of candidate causes (Fig. S2.8). Similarly, although there was weak evidence that confidence ratings were slightly more variable in vignettes with one candidate cause, there was no evidence for any such differences in vignettes with two, three, or four candidate causes (Fig. S2.9). All contrasts are reported in Table S2.1.

4.2.3. Graded causal judgments in conjunctive structures

We next tested whether causal judgments in conjunctive structures were graded according to the number of candidate causes. Indeed, causal judgments were lower and more variable in vignettes with four candidate causes than in vignettes with only one candidate cause both when the focal cause was normal and when it was abnormal. In line with the abnormal inflation effect, this decrease in causal judgments was stronger for normal causes than for abnormal causes (Fig. S2.10). However, there was no evidence that the increase in variability differed between normal and abnormal focal causes (Fig. S2.11). Finally, we tested whether confidence ratings in conjunctive structures were also affected by the number of candidate causes. Confidence ratings were slightly lower and more variable in vignettes with four candidate causes than in vignettes with one candidate cause when the focal cause was normal, but not when the focal cause was abnormal. Accordingly, this decrease in mean confidence and increase in variability of confidence ratings was stronger when the focal cause was normal than when it was abnormal (Figs. S2.12 & S2.13). We report all contrasts in Table S2.2.

4.2.4. Abnormal deflation generalizes to many causes

We then investigated the extent to which causal judgments in disjunctive structures were lower for abnormal causes than for normal causes (i.e., *abnormal deflation*). There was no evidence for abnormal deflation in vignettes with one or two candidate causes. However, there was strong evidence for abnormal deflation in vignettes with three candidate causes and evidence for weak abnormal deflation in vignettes with four candidate causes (Fig. S2.6). Causal judgments were also more variable for abnormal causes in vignettes with one, three, and four candidate causes, but not in vignettes with two candidate causes (Fig. S2.7). Finally, we tested whether abnormal deflation occurred in

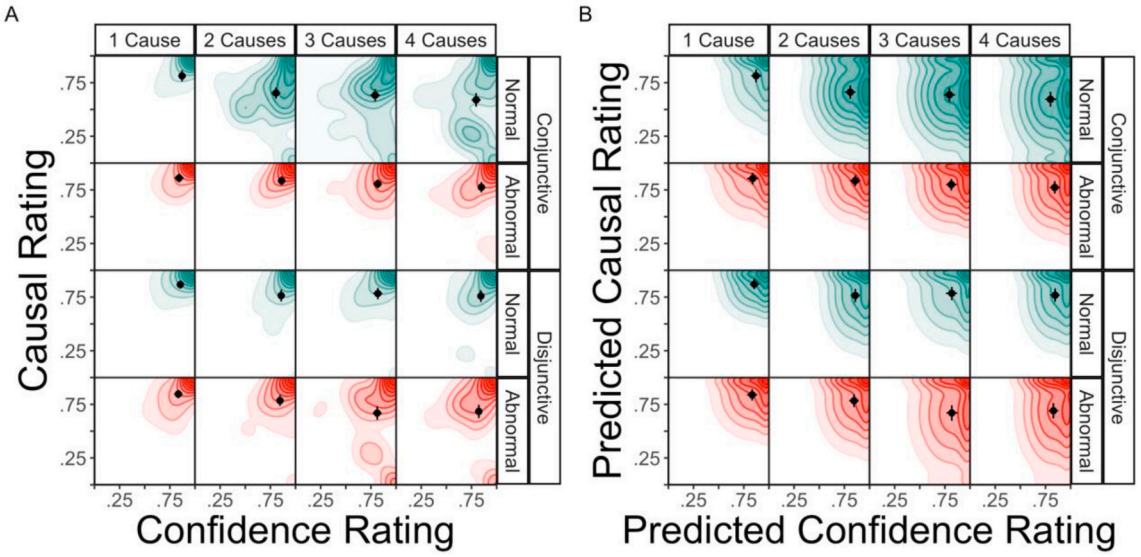

Fig. 4. Causal judgments, but not confidence, exhibit abnormal inflation, abnormal deflation, and gradation to the number of candidate causes. Density plots of distributions of causal judgments versus confidence for (*A*) the raw data from Experiment 2 and (*B*) the fitted model. Compared to causal judgments for normal candidate causes, judgments for abnormal candidate causes are higher in conjunctive structures (i.e., abnormal inflation) but lower in disjunctive structures (i.e., abnormal deflation). Points are means and the errorbars represent 95% CIs. Causal judgments decrease with the number of candidate causes in all conditions. In contrast, confidence is relatively unchanged by number of causes, causal structure, or normality.

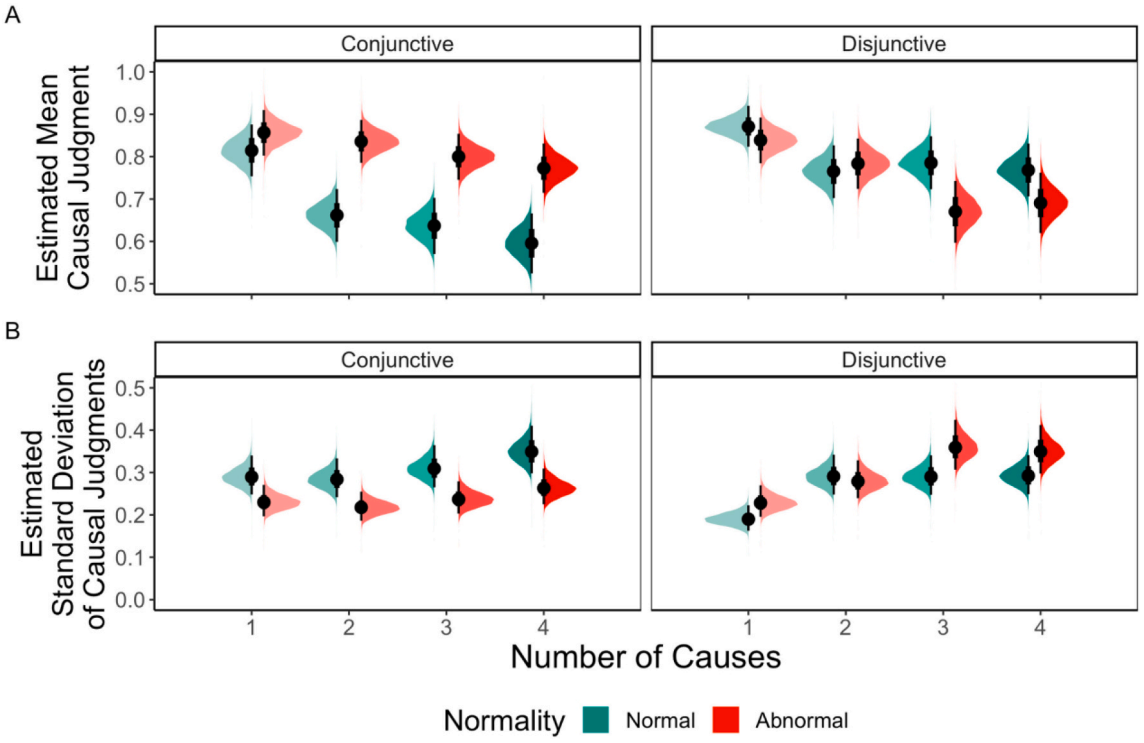

Fig. 5. Posterior mean and standard deviation of causal judgments exhibit abnormal inflation, abnormal deflation, and gradation to the number of candidate causes. Posterior estimates of mean causal judgment (*A*) and the standard deviation of causal judgments (*B*) from Experiment 2 as a function of causal structure, normality, and the number of candidate causes. Compared to causal judgments for normal candidate causes, judgments for abnormal candidate causes are higher and less variable in conjunctive structures (i.e., abnormal inflation) but lower and more variable in disjunctive structures (i.e., abnormal deflation). Causal judgments were also lower and more variable with an increasing number of candidate causes in all conditions. Dots represent medians, thick error bars represent 66% HDIs, and thin error bars represent 95% HDIs.

the mean or standard deviation of confidence ratings, but we found no evidence for such effects in vignettes with one, two, three, or four candidate causes (Figs. S2.8 & S2.9). We report all contrasts in Table S2.1.

4.2.5. *Graded causal judgments in disjunctive structures*

We then tested whether causal judgments in disjunctive structures were graded according to the number of candidate causes. As in conjunctive structures, there was strong evidence that causal judgments

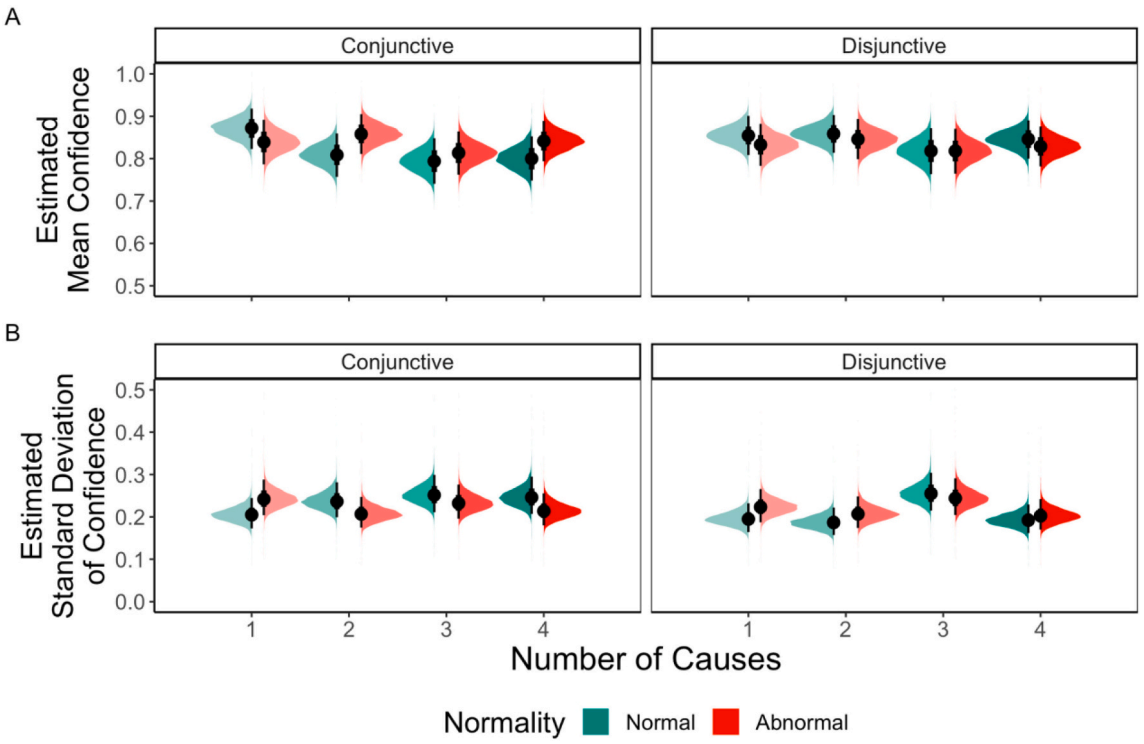

Fig. 6. Posterior mean and standard deviation of confidence in causal judgments was largely unaffected by normality, causal structure, and the number of candidate causes.

Posterior fits of mean confidence in causal judgment (*A*) and standard deviation of confidence in causal judgment (*B*) from Experiment 2 as a function of causal structure, normality, and the number of candidate causes. Overall, there were few detectable changes in the mean or standard deviation of confidence. Dots represent medians, thick error bars represent 66% HDIs, and thin error bars represent 95% HDIs.

were lower and more variable in vignettes with four candidate causes than in those with one candidate cause, both when the focal cause was normal and when it was abnormal. However, there was no evidence that the decrease in causal judgments or the increase in variability of judgments was different when the focal cause was abnormal compared to when it was normal (Figs. S2.10 & S2.11). Finally, there was no evidence for any such changes in the mean or standard deviation of confidence ratings when the focal cause was abnormal or when it was normal (Figs. S2.12 & S2.13). We report all contrasts in Table S2.2.

4.2.6. *Confidence does not explain effects on causal judgment*

Considering that confidence was largely unaffected by normality, causal structure, and the number of candidate causes, its ability to explain away the effects of these variables on causal judgment is limited. Nevertheless, to more directly evaluate the causal strength explanation, we additionally tested whether the effects of normality, causal structure, and number of candidate causes on causal judgment were mediated by confidence. A caution: an empirically unverifiable assumption of this mediation is that there were no hidden confounders between confidence and causal judgment that might cause an unexplained association between the two (Pearl, 2012). Additionally, to avoid running a separate mediation analysis for every coefficient in our model, we ran a single omnibus test for the presence of any direct effects of normality, causal structure, and number of candidate causes as a single treatment. Specifically, we performed a model comparison between two mediation models, each specified with the same dependent variables, vignette-level effects, and priors as the model reported above. The null model (Fig. S2.19B) was restricted to assume a full mediation of the effects of normality, causal structure, and the number of candidate causes on causal judgment through confidence. In other words, this model assumed that there were no direct effects of our manipulations on causal judgment. In contrast, our alternative model (Fig. S2.19A) predicted causal judgment in terms of both indirect and direct effects of our manipulations. This formulation loses precision in terms of delineating exactly which effects may or may not be mediated by confidence, but it is sufficient for our purposes: if there remains any direct effect of normality, causal structure, or the number of candidate causes after controlling for confidence, in the absence of any other explanation, we can say that the causal strength explanation must hold.

Comparing these two models, we found that the model with direct effects (ELPD = -70.9 (67.6), PSIS-LOOIC = 141.8 (135.2), Effective # of Parameters = 125.8 (7.7), Model Weight = .775) provided significantly better predictions of causal judgments than the model without direct effects (ΔELPD = -86.8 (24.1), ELPD = -157.7 (62.5), PSIS-LOOIC = 315.4 (125), Effective # of Parameters = 81.1 (5.8), Model Weight = .225), despite having a significantly larger number of parameters. Replicating results from Experiment 1, confidence was associated with causal judgments above and beyond normality, causal structure, and number of candidate causes as in Experiment 1, $\beta = .27$, 95% *HDI* = [.22, .32], *p-pd* < .001, 0% in ROPE. However, because it did not vary systematically between experimental conditions, confidence did not explain away the effects of our manipulations on causal judgment. Overall, we take these results to mean that confidence, normality, causal structure, and the number of candidate causes were largely independent predictors of causal judgment.

4.3. *Discussion*

The data from Experiment 2 suggest that effects on causal judgments due to normality, causal structure, and the number of candidate causes are not fully mediated by confidence. Thus, Experiment 2 provides support for the causal strength explanation of gradation in causal judgments: since the observed differences in causal judgments cannot be attributed to differences in confidence, in the absence of any other

explanation of these effects, we can conclude that people have graded concepts of causation.

We also conceptually replicated previous results showing that causal judgments for abnormal causes are higher than those for normal causes in conjunctive structures (i.e., *abnormal inflation*) but lower than those for normal causes in disjunctive structures (i.e., *abnormal deflation*). Notably, we also found parallel effects in the estimated variance of causal judgments: causal judgments of abnormal causes were less variable than those of normal causes in conjunctive structures, but more variable than those of normal causes in disjunctive structures. Our results thus extend those of Icard et al. (2017) to demonstrate that normality influences both mean causal judgment and the between-participant agreement on causal judgments.

Additionally, we found that causal judgments in all conditions became more graded (i.e., were lower and more variable) when more candidate causes were present. This result revealed a new perspective of abnormal inflation and deflation: normality affected the rate at which the strength of a candidate cause diminished with the addition of other candidate causes compared to when it preceded the effect alone (i.e., the difference between 1 and 4 candidate causes). In conjunctive structures, we found that causal judgments decreased with an increasing number of candidate causes for both normal and abnormal focal causes. Consistent with abnormal inflation, the decrease in causal judgment with increasing number of candidate causes was stronger for judgments of normal causes than for judgments of abnormal candidate causes, resulting in higher judgments of abnormal causes compared to judgments of normal causes. Likewise, in disjunctive structures, we found that causal judgments for normal and abnormal causes decreased as the number of candidate causes increased. However, because this decrease was similar for normal and abnormal causes, the abnormal deflation effect was weaker overall; abnormal deflation could only occur when the deflation of the abnormal cause was stronger than the deflation of the normal cause. This finding, along with vignette-specific differences (see Figs. S2.1–S2.4), may explain our failure to replicate abnormal deflation with two candidate causes. This conceptual replication failure is also consistent with other work demonstrating that this effect can be weak (Kirfel & Lagnado, 2018; Sytsma, 2019).

In stark contrast to the effects found in causal judgments, confidence in causal judgments was relatively unchanged by any of our manipulations. So, even though confidence was a reliable predictor of causal judgments in both of our experiments, it did not explain away the effects of normality, causal structure, and the number of candidate causes on causal judgments. In other words, the variation in causal judgments due to causal structure, normality, and the number of candidate causes was largely independent of the variation due to confidence. Although we leave it to future work to identify how participants keep track of uncertainty and assess the quality of their causal judgments, we speculate that several factors including individual differences in how participants interpret the vignettes and which counterfactuals participants consider may account for the relationship between confidence and causal judgment.

It is possible that confidence was unable to explain the effects of normality, causal structure, and the number of candidate causes in our mediation analysis simply because the relationship between confidence and causal judgments depends on the participants' discrete causal judgment (as we found in Experiment 1), which we did not measure in this experiment. To investigate this possibility, we performed the same analyses on the data from Experiment 1 that we report for Experiment 2 (see Supplementary Information section *Normality affects causal judgments and confidence*). While we found no effect of normality on confidence in Experiment 2 , we found a small but significant increase in confidence when the cause was abnormal in the conjunctive causal structure in Experiment 1 (Figs. S1.13-S1.14). We speculate that we did not find these small effects in Experiment 2 because confidence was at ceiling, and we suggest that future research should try to replicate these results in situations where people are less confident overall. However, confidence was still unable to fully mediate the effect of normality on causal judgments in Experiment 1, even when we allowed the effect of confidence to vary as a function of the discrete causal judgment. Overall, we can be sure that confidence alone cannot account for normality effects: in addition to the confidence explanation, we need the causal strength explanation of gradation in causal judgments.

5. General discussion

The current paper sought to address two major questions regarding singular causal judgments: are causal judgments graded, and if so, what explains this gradation? In the Preliminary Analyses, we found that causal judgments were largely multimodal, with the majority of responses either indicating a totally causal relationship or a totally non-causal relationship. There were, however, many responses in the center of the scale, indicating that causal judgments were to some extent graded.

Given that causal judgments were graded, we next asked why. In two experiments, we tested two competing (although not necessarily mutually exclusive) explanations. The causal strength explanation posits that causal judgments are graded according to gradation in participants' concepts of causation: weak causes should receive lower judgments than strong causes. The confidence explanation instead argues that causes are graded according to gradation in participants' certainty in their causal judgment: uncertain causes should receive lower judgments than certain causes. In line with the confidence explanation, we found in Experiment 1 that causal judgments were modulated by the degree of confidence in them. People tended to make graded causal judgments when they were uncertain, but tended to make more categorical judgments when they were certain. In Experiment 2, we replicated the result from Experiment 1 that confidence reliably predicted causal judgments. However, in line with the causal strength explanation, we found that confidence could not explain all gradation in causal judgments. Compared to judgments for normal candidate causes, causal judgments for abnormal candidate causes were higher and less variable in vignettes with conjunctive causal structures (i.e., abnormal inflation) but were lower and more variable in disjunctive causal structures (i.e., abnormal deflation). Additionally, causal judgments were lower and more variable when alternative candidate causes were present. But since confidence did not mediate these effects in causal judgments, we concluded that the remaining gradation must be attributable to a graded concept of causation.

Together, these results indicate support for both the confidence explanation and the causal strength explanation of gradation in causal judgment. In other words, people make graded causal judgments both when they think a cause is weak and also when they are uncertain about their causal judgment. Although work is needed to determine precisely why and when causal judgments are influenced by confidence, we have demonstrated that these effects are separable from more well-studied effects on causal judgment. This is good news for theories of causal judgment that rely on the causal strength explanation: these theories do not need to account for the effects of confidence on causal judgment to be useful in explaining other effects. That is, there is no need for major revisions in how we think about causal judgments. Nevertheless, we think our results have important implications for these theories, which we outline below.

5.1. Implications for theories of causal judgment

Overall, our results have clear implications for dependence theories of causal judgment (i.e., those positing that causes make a difference to their effects). The remaining dependence theories invoking non-graded concepts of causation, including the mental model theory of causal judgments, must be extended to account for the observed gradation in these judgments (Bello et al., 2018; Goldvarg & Johnson-Laird, 2001; Khemlani et al., 2014; Khemlani et al., 2018). Fortunately, the mental model theory can be extended in several ways to account for gradation,

including weighting causal judgments by the strength of evidence for them and the type of underlying causal relation that they reflect (Johnson-Laird & Khemlani, 2017). On the other hand, dependence accounts that invoke graded concepts of causation require additional assumptions to explain why there is often such little gradation in causal judgments (Cheng, 1997; Gerstenberg et al., 2017; Icard et al., 2017; Quillien, 2020). One possibility is that these models could appeal to confidence to explain non-graded causal judgments: if participants are sufficiently confident (as is highly likely with simple vignette stimuli), then their causal judgments will be less graded. Alternatively, these models could be augmented with a multimodal or censored data-generating process to predict the prevalence of extreme causal judgments. Yet another possibility is that they could appeal to individual differences in the information participants consider when making causal judgments (e.g., Gerstenberg et al., 2021).

In contrast to dependence theories, process theories of causal judgment (including the force dynamics theory) argue that people judge whether an event is causal by asking whether there was a particular exchange of force between a cause and effect (Wolff, 2007). While it is well known that process theories fail to account for other effects on causal judgments (Bernstein, 2014; Danks, 2017; Henne, Niemi, et al., 2019), they may be consistent with the main findings in this paper. Strictly interpreted, the force dynamics theory does not model concepts of causation as graded: a process is causal if and only if there was an appropriate transfer of force between the cause and effect, and the qualitative direction of this transfer of force determines the type of causal relation between the cause and effect. But this model could easily be extended to account for gradation simply by appealing to the quantitative magnitude, and not just the qualitative direction, of the force on the effect. Put simply, causes that transfer more force to an effect are represented as more causal than those that transfer less force. Just like the mental model theory, the force dynamic theory could also explain gradation in causal judgments by weighting causal judgments according to whether the particular transfer of force indicates a causal, enabling, or prevention relationship (Wolff et al., 2010). It seems more difficult to see why confidence should be related to causal judgments from this perspective, but the force dynamics model can account for this finding as well. In particular, Wolff acknowledges that uncertainty about the precise magnitude or direction of forces creates uncertainty about the overall configuration of forces, which propagates along causal chains (Wolff, 2007). So, under the force dynamics model, people may give graded causal judgments when they are uncertain about the magnitude or direction of relevant forces.

Social cognition accounts of causal judgment, such as the culpable control model, are also consistent with our findings (Alicke, 2000; Alicke et al., 2011). Under this account, people make causal judgments to confirm ascriptions of blame to negatively-evaluated agents. To the extent that blame ascriptions are graded (which is a worthy topic of investigation in its own right), this account naturally allows for causal judgments to be graded in terms of how blameworthy an agent is. Though uncertainty is not explicitly included in their model, Alicke et al. (2011) predict and present data implying that confidence is relevant to causal judgment. Specifically, they demonstrate that the extent to which blame ascriptions influence causal judgments is moderated by the presumed ambiguity of a causal structure: when a causal structure is ambiguous, people's causal judgments are more influenced by blame ascriptions than when the causal structure is unambiguous (Alicke et al., 2011: Experiment 3). Alicke et al. (2011) did not collect confidence ratings, however, so more evidence is required to demonstrate that uncertainty moderates the relationship between blame ascriptions and causal judgments.

5.2. Recommendations for future research

In addition to having implications for individual theories of causal judgment, our results indicate two broad areas of improvement overall: taking gradation and uncertainty seriously. First, if we are interested in causal judgment as a graded phenomenon, it does not suffice to account only for changes in mean causal judgment. To explain precisely when, how, and why causal judgments are graded, we must shift the focus of our explanations to distributions of individual causal judgments. For instance, we have seen in the Preliminary Analyses that in some cases, a graded mean causal judgment is reflective of between-participant agreement on a graded judgment, but in other cases, a graded mean judgment simply reflects between-participant disagreement overall (e.g., Henne, Niemi, et al., 2019; Icard et al., 2017). Though more work is needed to tease apart these explanations, it is possible that these differences in the extent and form of gradation in causal judgments could arise from individual differences in the information considered when making a causal judgment (Gerstenberg et al., 2021; Osman & Shanks, 2005), differences in the stimuli or measurement scales used, differences in the amount of uncertainty, or other contextual factors. But because the mean alone cannot differentiate between these cases (i.e., of participant-level agreement and disagreement), distinguishing them requires extending both the methodological tools used to study causal judgments and theories of causal judgment themselves.

Descriptively, generative models of causal judgments may afford more informative interpretations of how individuals, and not just groups, make causal judgments, and may even serve to advance cognitive models of these judgments (Haines et al., 2020; Kennedy et al., 2019; Schad et al., 2021). Although we acknowledge that other distributions may provide even better descriptions of causal judgments, unequal-variance Gaussian models outperformed standard equal-variance Gaussian models in both of our experiments. At the very least, we recommend that researchers regularly plot histograms of their data and compare different statistical models of their data to avoid making unmet assumptions and erroneous inferences. In terms of theory, models of causal judgment can and should aim to characterize distributions of causal judgments, and not simply patterns in mean causal judgment. Following recent work focusing on individual differences between participants in causal judgments (Gerstenberg et al., 2021; Godfrey-Smith, 2009; Osman & Shanks, 2005), we think that models of causal judgment that are already successful at the group-level could greatly benefit from extending their explanations to variation between participants and even variation within participants at the trial-level. For example, Gerstenberg et al. (2021) demonstrated that differences in the extent to which individuals consider different aspects of causation can account for differences in their judgments of causal responsibility. Accounting for these kinds of differences in agreement and disagreement between participants is not only an important test for models of causal judgment, but also it is an important area of investigation in its own right.

Second, although research in causal cognition has largely overlooked the role of confidence in causal judgment, there is potentially much to gain from developing an explanation of this relationship. In the domain of general causal judgments, the causal support model parsimoniously explains a wide array of findings on causal judgment, causal structure induction, and causal explanation in terms of participants' certainty that a link exists between cause and effect, as opposed to their inferred strength of a causal relationship (Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011; Lombrozo, 2007; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Meder, Mayrhofer, & Waldmann, 2014; Tenenbaum & Griffiths, 2001). Even for the simple vignettes used in our experiments, we found that confidence is also an important predictor of singular causal judgments. For more complex problems with larger causal structures, unobserved variables, or perceptual uncertainty, we can only expect confidence to play an even larger role. Thus, developing accounts of how people keep track of uncertainty will become increasingly imperative to developing accounts of causal judgments themselves. Finally, although confidence could not explain our observed differences in causal judgments due to normality, causal structure, or the number of candidate causes, it very well may explain differences in causal

judgments due to other experimental manipulations. To ensure that confidence does not confound results in future work, it is necessary to collect confidence ratings and to control for confidence. Given the relevance of uncertainty to causal judgment, we think this is a small price to pay to meaningfully interpret causal judgments in terms of peoples' causal concepts.

5.3. Conclusion

In sum, we examined the presence and explanation of gradation in singular causal judgments. In the Preliminary Analyses, we found that although people usually make categorical causal judgments even when given a continuous scale, they also often make graded causal judgments. In Experiment 1, we found that gradation in causal judgments is modulated by confidence. People tended to make graded causal judgments when they were less confident and they tended to make more categorical causal judgments when they were more confident. In Experiment 2, we found that causal judgments for abnormal causes were higher than for normal causes in conjunctive causal structures (i.e., abnormal inflation) but lower than for normal causes in disjunctive structures (i.e., abnormal deflation) and that causal judgments were more graded when more candidate causes were present. Even though confidence independently predicted causal judgments, these effects could not be explained by confidence because confidence was relatively constant in all conditions. In sum, we found that causal judgments have at least two independent sources of gradation: gradation in people's concepts of causation, and gradation in their certainty about their causal judgment.

Open practices statement

The data, analysis code, and materials for all experiments are available on the Open Science Framework https://osf.io/dwjpt/ and none of the experiments were preregistered.

Declaration of Competing Interest

None.

Acknowledgments

This research was supported by a grant from the Office of Naval Research (N00014-17-1-2603) to FDB. We would also like to thank Maria Khoudary, Will Bridewell, Andrew Lovett, Sangeet Khemlani, and two anonymous reviewers for providing valuable feedback on various stages of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105036.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556.
Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy, 108*(12), 670–696.
Bello, P., & Khemlani, S. (2015). A model-based theory of omissive causation. In *Proceedings of the 37th annual conference of the cognitive science society*.
Bello, P., Lovett, A. M., Briggs, G., & O'Neill, K. (2018). An attention-driven computational model of human causal reasoning. In *Proceedings of the 40th annual meeting of the cognitive science society*.
Bernstein, S. (2014). Omissions as possibilities. *Philosophical Studies, 167*(1), 1–23.
Bernstein, S. (2017). Causal proportions and moral responsibility. *Oxford Studies in Agency and Responsibility, 4*, 165–182.
Bürkner, P. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.
Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology, 67*, 135–157.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367.
Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*(4), 545.
Collins, D. J., & Shanks, D. R. (2006). Short article conformity to the power PC theory of causal induction depends on the type of probe question. *The Quarterly Journal of Experimental Psychology, 59*(2), 225–232.
Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.
Danks, D. (2013). Functions and cognitive bases for the concept of actual causation. *Erkenntnis, 78*(1), 111–128.
Danks, D. (2017). Singular causation. In *The Oxford handbook of causal reasoning* (pp. 201–215).
Dowe, P. (1992). Wesley Salmon's process theory of causality and the conserved quantity theory. *Philosophy of Science, 59*(2), 195–216.
Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2.
Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 523–528). Cognitive Science Society.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*.
Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General, 149*(3), 599.
Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731–1744.
Godfrey-Smith, P. (2009). Causal pluralism. In *The Oxford handbook of causation* (pp. 326–337).
Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*(4), 565–610.
Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 334–384.
Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology, 11*, 1069.
Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., & Turner, B. (2020). *Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox* (PsyArXiv).
Hall, N. (2004). Two concepts of causation. In *Causation and counterfactuals* (pp. 225–276).
Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science, 66*(2), 413–457.
Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. OUP Oxford.
Henne, P., Bello, P., Khemlani, S., & De Brigard, F. (2019). Norms and the meaning of omissive enabling conditions. In *Proceedings of the 41st annual conference of the cognitive science society* (p. 41).
Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition, 212*, Article 104708.
Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition, 190*, 157–164.
Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science, 45*(1), Article e12931.
Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy, 95*(2), 270–283.
Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review, 93*(1), 75.
Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science, 79*(5), 942–951.
Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy, 106*(11), 587–612.
Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology, 62*, 135–163.
Hume, D. (2000). *An enquiry concerning human understanding: A critical edition* (Vol. 3). Oxford University Press (Original work published 1748).
Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition, 161*, 80–93.
Jackson, E. G. (2020). The relationship between belief and credence. *Philosophy Compass, 15*(6), Article e12668.
Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and applied, 79*(1), 1.
Johnson-Laird, P. N., & Khemlani, S. (2017). Mental models and causation. In *Oxford handbook of causal reasoning* (pp. 1–42).
Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136.
Kaiserman, A. (2016). Causal contribution. In *, 116(3). Proceedings of the Aristotelian society* (pp. 387–394). Oxford University Press.
Kaiserman, A. (2018). 'More of a cause': recent work on degrees of causation and responsibility. *Philosophy Compass, 13*(7), Article e12498.
Kennedy, L., Simpson, D., & Gelman, A. (2019). The experiment is just as important as the likelihood in understanding the prior: A cautionary note on robust cognitive modeling. *Computational Brain & Behavior, 2*(3), 210–217.
Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & Cognition, 46*(8), 1344–1359.
Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience, 8*, 849.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition, 212*, Article 104721.
Kirfel, L., & Lagnado, D. A. (2018). Statistical norm effects in causal cognition. In *Proceedings of the 40th annual conference of the cognitive science society*.
Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology, 2*, 441–448.
Kolvoort, I. R., Davis, Z. J., van Maanen, L., & Rehder, B. (2021). Variability in causal judgments. In *, 43(43). Proceedings of the annual meeting of the cognitive science society*.
Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science, 43*(11), Article e12792.
Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition, 137*, 196–209.
Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312.
Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178–206.
Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*(3), 754–770.
Lagnado, D. A., Gerstenberg, T., & Zultan, R. I. (2013). Causal responsibility and counterfactuals. *Cognitive science, 37*(6), 1036–1073.
Lewis, D. (1973). Causation. *The Journal of Philosophy, 70*(17), 556–567.
Liljeholm, M., & Cheng, P. W. (2009). The influence of virtual sample size on confidence and causal-strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 157.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232–257.
Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*(4), 303–332.
Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*(4), 955.
Makowski, D., Ben-Shachar, M. S., Chen, S. H., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology, 10*, 2767.
Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General, 132*(3), 419.
Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of personality and social psychology, 71*(3), 450.
McEleney, A., & Byrne, R. M. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning, 12*(2), 235–255.
McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 123*(1/2), 125–148.
Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*(3), 277.
Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLoS One, 14*(8).
Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). *Causal judgments approximate the effectiveness of future interventions* (PsyArxiv).
N'gbala, A., & Branscombe, N. R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology, 31*(2), 139–162.
Osman, M., & Shanks, D. R. (2005). Individual differences in causal learning and decision making. *Acta Psychologica, 120*(1), 93–112.
Pearl, J. (2009). *Causality*. Cambridge University Press.
Pearl, J. (2012). The causal mediation formula—A guide to the assessment of pathways and mechanisms. *Prevention Science, 13*(4), 426–436.
Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology Section A, 56*(6), 977–1007.
Quillien, T. (2020). When do we think that X caused Y? *Cognition, 205*, Article 104410.
Quillien, T., & Barlev, M. (2021). *Causal judgment in the wild: evidence from the 2020 US presidential election* (PsyArXiv).
Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308.
Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science, 61*(2), 297–312.
Sartorio, C. (2020). More of a Cause? *Journal of Applied Philosophy, 37*(3), 346–363.
Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods, 26*(1), 103.
Schlottmann, A., & Anderson, N. H. (1993). An information integration approach to phenomenal causality. *Memory & Cognition, 21*(6), 785–801.
Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation, 18*(2), 147–166.
Shou, Y., & Smithson, M. (2015). Effects of question formats on causal judgments and model evaluation. *Frontiers in Psychology, 6*, 467.
Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*(4), 323.
Spellman, B. A., & Ndiaye, D. G. (2007). On the relation between counterfactual and causal reasoning. *Behavioral and Brain Sciences, 30*(5–6), 466–467.
Sprenger, J. (2018). Foundations of a probabilistic theory of causal strength. *Philosophical Review, 127*(3), 371–398.
Sytsma, J. (2019). Structure and norms: Investigating the pattern of effects for causal attributions. Preprint http://philsci-archive.pitt.edu/16626/.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In *Advances in neural information processing systems* (pp. 59–65).
Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science, 42*(4), 1265–1296.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432.
Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*(1), 82.
Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General, 139*(2), 191.
Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review, 115*(1), 1–50.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.