Causality, Normality, and Sampling Propensity

Thomas F. Icard (icard@stanford.edu)
Philosophy & Symbolic Systems, Stanford University

Joshua Knobe (joshua.knobe@yale.edu)
Cognitive Science & Philosophy, Yale University

Abstract

We offer an account of the role of normality—both statistical and prescriptive—in judgments of actual causation. Using only standard tools from the literature on causal cognition, we argue that the phenomenon can be explained simply on the assumption that people stochastically sample (counterfactual) scenarios in a way that reflects normality. We show that a formalization of this idea, giving rise to a novel measure of causal strength, can account for some of the most puzzling qualitative patterns uncovered in recent experimental work.

Introduction

Judgments of *actual causation*—concerning the extent to which a given event or factor *caused* some outcome on a particular occasion—have been at the center of attention in work on causal cognition. One intriguing phenomenon that has long been recognized is that people's judgments of actual causation can be influenced by the degree to which they regard certain events as *normal*. In recent years, this effect has been explored both in experimental studies (Kominsky et al., 2015; Phillips et al., 2015) and in formal models (e.g., Halpern and Hitchcock 2015).

In this paper we propose a novel explanation of the role of normality in causal cognition by appeal to the idea that many cognitive processes—including those underlying causal judgments—can be understood as involving *probabilistic sampling* from some underlying distribution. In short, causal strength is assessed in part by stochastically generating alternative ("counterfactual") scenarios and using these scenarios to determine the extent to which some event is causally relevant to a given outcome. This hypothesis, together with a further assumption—that the distributions from which these samples are drawn directly reflect prescriptive normality (as well as statistical factors)—forms the core of our explanation.

To explore this possibility, we begin by summarizing the effects to be explained and describing at an informal level how these effects could be explained in terms of sampling. Then we turn to the details of a natural sample-based algorithm and show that the *measure of actual causal strength* corresponding to this algorithm, unlike other causal strength measures offered in the literature, would generate precisely the effects observed in existing studies.

Three Effects of Normality on Actual Causation Judgments

To begin with, we need to distinguish two kinds of norms. First, there are purely *statistical norms*. For example, it is a statistical fact that winter months in Oregon tend to be cloudy and overcast, so if Oregon ever had a sunny winter, that weather could be said to be violating a statistical norm. Second, there are *prescriptive norms*. These norms are constituted not by purely statistical tendencies but by the way things ought to be or are supposed to be. Suppose we believe that the police ought to accord criminal defendants certain rights. Even if we do not believe that the police actually do tend to accord defendants these rights, we might think that failing to do so is a violation of a prescriptive norm.

A question then arises as to which of these two types of norms impacts people's judgments of actual causation. As we will see below, existing research suggests that actual causation judgments are influenced by both kinds of norms. More strikingly, the impact of these two kinds of norms shows precisely the same pattern. As a result, researchers have suggested that it might be helpful to posit a single undifferentiated notion of normality that integrates both statistical and prescriptive considerations (Halpern and Hitchcock, 2015; Kominsky et al., 2015). On this approach, an event counts as "abnormal" to the extent that it either violates a statistical norm or violates a prescriptive norm, and as "normal" to the extent that it follows both of these types of norms. Difficult questions arise about precisely how statistical and prescriptive considerations are integrated into an undifferentiated notion, but we will not be resolving those questions here. Instead, we focus on three ways in which normality impacts people's intuitions about actual causation.

First Effect: Abnormal Selection

We will eventually be introducing a formal framework to describe this effect more precisely, but for the moment, we can offer the following rough characterization:

> *In cases where an outcome depends on a causal factor C, people will be more inclined to say that C caused the outcome when they regard C as abnormal than when they regard C as normal.*

This basic effect appears to arise both for statistical norms and for prescriptive norms.

First, it has been known for decades that actual causation judgments can be influenced by statistical norms (Hilton and Slugoski, 1986). Suppose that a person leaves a lit match on the ground and thereby starts a forest fire. In such a case, the fire would not have begun if there had been no oxygen in the atmosphere, and yet we would not ordinarily say that the oxygen caused the fire. Why is this? The answer appears to be that it is so normal for the atmosphere to contain oxygen. (Our intuitions would be very different if matches were struck

on a regular basis but there was never a fire except on the very rare occasions when oxygen was present.)

Strikingly, this same effect arises for prescriptive norms. Consider the following case:

The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly e-mailed them reminders that only administrators are allowed to take the pens.

On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk.

Faced with this case, participants tend to say that the professor caused the problem (Knobe and Fraser, 2008; Phillips et al., 2015). But now suppose that we change the first paragraph of the case in such a way as to make the professor's action not violate a prescriptive norm:

The receptionist in the philosophy department keeps her desk stocked with pens. Both the administrative assistants and the faculty members are allowed to take the pens, and both the administrative assistants and the faculty members typically do take the pens. The receptionist has repeatedly e-mailed them reminders that both administrators and professors are allowed to take the pens.

Faced with this latter version, participants are significantly less inclined to say that the professor caused the problem (Phillips et al., 2015). Yet the two cases do not appear to differ from the perspective of purely statistical normality; the difference is rather in the degree to which the agent violates a prescriptive norm. The result thereby suggests that prescriptive norms impact causal judgments.

This phenomenon has been explored in a wide range of studies (Cushman et al., 2008; Roxborough and Cumby, 2009; Samland et al., 2016), and the results strongly suggest that the effect really does involve prescriptive considerations and cannot be reduced to a matter of purely statistical norms. First, one can explicitly pit the prescriptive against the statistical. In one study, participants were told that administrative assistants were allowed to take pens and faculty members were not (a prescriptive norm) but that in actual fact administrative never did take pens while faculty members always did (a statistical norm). People's judgments ended up being affected more by the prescriptive than by the statistical, with participants tending on the whole to say that the administrative assistant did not cause the problem but the faculty member did (Roxborough and Cumby, 2009). Second, one can look at cases in which different people have different prescriptive judgments. For example, one paper looked at controversial political issues (abortion, euthanasia) and found that people who had opposing moral judgments about these issues arrived at correspondingly opposing causal judgments about people who performed the relevant actions (Cushman et al., 2008).

At this point, it is widely agreed that prescriptive considerations do indeed impact people's causal judgments. The remaining questions are about how to explain this. Existing accounts invoke everything from conversational pragmatics to motivational bias (see Livengood and Rose 2016 for an overview of the literature). Rather than introducing some additional component above whatever is used to account for causal judgments, our account will derive the normality effects as a straightforward consequence of the way counterfactuals are generated to determine causal strength.

Second Effect: Supersession

Supersession is an effect whereby the normality of one factor can actually influence the degree to which other factors are regarded as causes. The effect can be characterized roughly as follows:

Consider cases in which an outcome depends on two different factors C and A, such that the outcome will only occur if both C and A occur. Then people will be less inclined to say that C caused the outcome if A is abnormal than if A is normal.

In other words, it is not just that a given factor is regarded as more causal when it is abnormal; a factor will be also be regarded as more causal when other factors are normal. This effect too arises for both statistical and prescriptive norms. Turning first to statistical norms, consider the following:

Alex is playing a board game. On every turn of the game, two six-sided dice are rolled and a coin is flipped. Alex will either win or lose the game on his next turn.

Alex will only win the game if the total of his dice roll is greater than 2 AND the coin comes up heads. It is very likely that he will roll higher than 2, and the coin has equal odds of coming up heads or tails.

Alex flips the coin and rolls his dice at exactly the same time. The coin comes up heads, and he rolls a 12, so as expected, he rolled greater than 2. Alex wins the game.

Now contrast that with a case in which the second paragraph is slightly modified:

Alex is only win the game if the total of his dice roll is greater than 11 AND the coin comes up heads. It is very unlikely that he will roll higher than 11, but the coin has equal odds of coming up heads or tails.

The difference between these two cases is solely in the normality of the dice roll. (The success of the dice roll is statistically normal in the first case, statistically abnormal in the second.) Yet this difference actually leads to a change in the degree to which people regard the coin flip as a cause. Participants were significantly less inclined to say that Alex won because of the coin flip when the dice roll was abnormal than when it was normal (Kominsky et al., 2015).

This same effect then arises for prescriptive norms. In one study, participants were asked to imagine a motion detector that goes off whenever two people are in the room at the same

time. Suzy and Billy enter the room at the same time, and the motion detector goes off. In one condition, Billy is supposed to be in the room, while in the other condition, he is specifically not supposed to be in the room. Suzy was judged to be significantly less a cause of the detector going off when Billy violated the prescriptive norm than when he acted in accordance with the prescriptive norm (Kominsky et al., 2015).

Third Effect: No Supersession with Disjunction

The supersession effect described above arises in cases where the causal structure is conjunctive. For disjunctive structures we find a quite different pattern:

> *Consider cases in which an outcome depends on two different factors C and A, such that it will only occur if either C or A occurs. Then people are just as inclined to say that C caused the outcome when A is abnormal as they are when A is normal.*

Existing sudies have put this claim to the test by comparing disjunctive cases to conjunctive cases and looking for an interaction whereby manipulations of normality do not have the impact in disjunctive cases that they do in conjunctive ones. This interaction arises both for statistical norms and for prescriptive norms (Kominsky et al., 2015).

For statistical norms, we can see the effect by looking at the case of the coin flip and dice roll described above. One can simply modify the rules described in that case so that Alex wins if he succeeds either on the coin flip or on the dice roll. When the rules are changed in this way, the supersession effect disappears. Participants are just as inclined to see the coin flip as causal in the case where the dice roll is abnormal as they are in the case where the dice roll is normal.

Precisely the same result then arises for the prescriptive norm case with the motion detector. When participants are told that the motion detector will go off if at least one person is in the room, the supersession effect again disappears. Participants are just as inclined to see Suzy as the cause when Billy's act violates a prescriptive norm as when it does not.

Summary

Across three different effects, prescriptive norms appear to be having the same impact as statistical norms. If there had only been an impact of statistical norms, one obvious approach would have been to explain that impact in terms of something specific to the statistical case in particular—though even that would be difficult (see below)—but given that these effects appear to arise for both kinds of norms, it seems that we need a unified explanation that can be applied to both.

Sampling Propensities

The idea that various mental operations can be described in terms of probabilistic sampling processes has generated much excitement over the past several years (see, e.g., Griffiths et al. 2012; Icard 2016 for overviews). Our focus in this paper is on the role of sampling in causal cognition. We follow a long line of work that proposes, in judging whether some event C caused E, we must consider various counterfactual scenarios involving C and E (Lewis, 1973). In short, the sampling hypothesis in the causal domain proposes that these counterfactual scenarios will be selected and evaluated stochastically by a sampling-like process.

Different researchers have spelled out the counterfactual approach to causality in quite different ways. The most important point to emphasize is one that is common to many such implementations. This is that, in assessing any given causal claim, a wide variety of different counterfactual scenarios will be relevant. It is generally agreed that, at a minimum, one must check whether the cause C was in some sense *necessary* for effect E to happen. Necessity is usually thought to correspond to a counterfactual claim of the form: Had C not occurred, E also would not have occurred. Clearly there are many possible ways an event C might not occur, and theories of causation will often attempt to specify exactly which such non-C scenarios are relevant to the causal claim.

Dual to necessity, a number of researchers have proposed that judgments of actual causation also involve a notion of *sufficiency* (Woodward, 2006; Lombrozo, 2010): Given that C in fact occurred, the outcome E still would have occurred even if background conditions had been slightly different. To the extent that this counterfactual sufficiency claim does not hold, that generally counts against the causal strength of C on outcome E. Again, there are clearly many ways the background conditions might have been different.

Thus, independent of any particular proposal about exactly which counterfactuals are relevant, it is clear that there may in general be very many relevant counterfactual scenarios to consider, certainly more than any person could plausibly assess in real time. We propose that people solve this intractability problem by probabilistically *sampling* possible states of the world (for related ideas see Lucas and Kemp 2015, Gerstenberg et al. 2014). It is not as though there is some fixed set of counterfactual scenarios one must evaluate; rather, on any given occasion one will stochastically consider various alternative sequences of events, and thereupon make a judgment of causal strength.

Given therefore that a person will generate counterfactual scenarios with different probabilities, and will evaluate those scenarios probabilistically, the interesting question is how to understand these *sampling propensities*. In as far as sampling algorithms are typically used to approximate probabilistic calculations, it would make sense for these counterfactual sampling propensities to reflect the perceived environmental statistics. Indeed, one can interpret much of the recent work on sampling in cognition, including in causal cognition, as proposing that some species of subjective probability is indirectly represented by sampling propensities. However, the distinctive feature of our proposal in this paper is that sampling propensities are also proportional to *prescriptive* normality. Thus for instance, in the detector scenario when judging whether Suzie caused the alarm, if Billy is prohibited from being in the room, people will be more likely to con-

sider counterfactual situations in which he did not enter.

It bears emphasis that sampling propensities, on this view, need not (only) encode the subject's uncertainty. It has already been suggested recently that people's sampling distributions in some contexts might come apart from reasonable judgments of probability or likelihood (see Lieder et al. 2014; Icard 2016), e.g., in a way that favors practical rationality (making the right choice) at the expense of theoretical rationality (having true or accurate beliefs). But some aspects of sampling propensity might be altogether divorced from probabilistic judgment. Consider, e.g., what might happen if we are observing an agent going through a series of steps in an attempt to solve a difficult puzzle. As we observe her taking each step, we consider other possible steps she might have taken. In such a case, we might be especially likely to consider other possible steps that would have been *good* ways of solving the puzzle (Kahneman and Miller, 1986). We consider these possibilities not because we explicitly represent the agent as having a high probability of taking them but, perhaps, because we are interested in finding the solution and are drawn to consider the best ways of accomplishing that task.

Similar remarks apply to the case we originally introduced to illustrate the concept of a prescriptive norm. Suppose we believe that the police should accord criminal defendants certain rights but that, as a purely statistical matter, they almost never do accord those rights. Now suppose we are observing a case in which the police fail to accord a defendant her rights. In such a case, we might consider possibilities in which the police do accord the defendant these rights. We would be drawn to consider those possibilities not because we explicitly regard them as probable but, perhaps, rather because we see them as instantiating a moral ideal.

Existing research provides some support for this hypothesis. When participants are given a vignette and asked to provide a counterfactual, they are more likely to mention possibilities they regard as statistically frequent (Kahneman and Miller, 1986). Furthermore, they are also more likely to mention possibilities they regard as prescriptively good (McCloy and Byrne, 2000). In addition, when participants are given a counterfactual and asked to rate the degree to which it is relevant or worth considering, they are more inclined to rate a possibility as relevant to the extent that it conforms to prescriptive norms (Phillips et al., 2015). These findings provide at least some initial evidence in favor of the claim that people are drawn to consider possibilities that do not violate prescriptive norms. The idea that such inclinations would similarly play into the hypothesized sampling propensities underlying causal and other judgments does not seem implausible.

Significantly, the mix of statistical and prescriptive norms that we find in people's actual causation judgments has also arisen in a number of other areas. For example, one finds a similar effect in people's intuitions about intentional action and about freedom (Phillips et al., 2015). Existing attempts to explain these effects have suggested that they might be due in some way to people's tendency to regard possibilities as more relevant when those possibilities accord with norms, though this basic approach has been spelled out within a number of different formal frameworks (e.g., Knobe and Szabó 2008; Halpern and Hitchcock 2015). If all of these phenomena can be elegantly explained in terms of sampling, this would provide strong evidence in favor of the present hypothesis.

Ultimately, however, the real test of this hypothesis is whether, when combined with further assumptions, it can accurately predict the patterns of people's causal judgments.

The Account

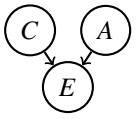
We formalize the notions of sufficiency and necessity in a straightforward way using Bayes nets and causal interventions (Pearl 2009), where we assume that the underlying probabilities reflect normality judgments as discussed in the previous section. That will affect not only the sufficiency/necessity judgments themselves, but also the probabilities with which sufficiency and necessity are assessed, effectively giving a weighting between them. This plays an important role in our account of abnormal selection.

Intervening on a Bayes net $\mathcal{N} = \langle \mathcal{G}, P \rangle$ involves setting some variable X to a specific value x. This gives rise to a new mutated Bayes net $\mathcal{N}_{X=x} = \langle \mathcal{G}_{X=x}, P_{X=x} \rangle$, where in the graph $\mathcal{G}_{X=x}$ we cut all links to the node representing X, so that it has no parents, and in $P_{X=x}$ we have $P_{X=x}(X = x) = 1$, leaving everything else the same. We can then use this to infer what would have happened under various counterfactual suppositions, including in cases where we nonetheless want to keep some observations from the actual situation fixed. That is, we can also use $P_{X=x}(Y \mid \vec{Z} = \vec{z})$ to determine counterfactual probabilities of another variable Y, holding fixed that we actually observed $\vec{Z} = \vec{z}$, for example.

We adopt standard notation for interventions and conditionalization. Given a network $\mathcal{N} = \langle \mathcal{G}, P \rangle$ we will write $P\big(Y = y \mid do(X = x)\big)$ for $P_{X=x}(Y = y)$ and similarly for intervention with observations, $P\big(Y = y \mid do(X = x), \vec{Z} = \vec{z}\big) = P_{X=x}(Y = y \mid \vec{Z} = \vec{z})$. We will also use somewhat nonstandard notation for negative intervention: $P\big(Y \mid do(X \neq x)\big) = P_{X \neq x}(Y)$, where $P_{X \neq x}$ is just like P, except that $P_{X \neq x}(X = x) = 0$ and $P_{X \neq x}(X = x')$ is the renormalized probability of $X = x'$, i.e., $\frac{1}{Z} P(X = x')$, with $Z = \sum_{x' \neq x} P(X = x')$.

Desiderata

All of our motivating examples involve a simple 3-node graph (known in the literature as an "unshielded collider" structure):



Assuming the random variables A, C, E are all binary—taking on values 0 and 1—and given a distribution P that factors over this graph, we are interested in two special cases:

CONJUNCTIVE: $P(E \mid C, A) = \min(C, A)$

DISJUNCTIVE: $P(E \mid C, A) = \max(C, A)$

In other words, the conjunctive version has E on (value 1) if both C and A are on, off (value 0) otherwise. This kind of model would describe the scenario with the pens, for example: the receptionist has a problem ($E = 1$) just in case both the administrator takes a pen ($C = 1$) and the professor takes a pen ($A = 1$). By contrast, the disjunctive version has E on if at least one of C or A is on. This describes the disjunctive scenarios: e.g., the motion detector goes off just in case either Billy enters the room or Suzy enters the room.

To account for the effects we need a definition of *causal strength*. Suppose we have a functional $\kappa_P(C,E)$ that measures the strength of cause $C = 1$ on effect $E = 1$ under distribution P. We can then rewrite our desiderata in a slightly more formal manner. Suppose that P_1 and P_2 are two distributions such that $P_1(C) = P_2(C)$, but that $P_1(A) > P_2(A)$. Then depending on whether P_1 and P_2 are conjunctive or disjunctive, we should expect different patterns:

ABNORMAL SELECTION: In the conjunctive case, $\kappa_{P_1}(A,E) < \kappa_{P_2}(A,E)$.
SUPERSESSION: Again, in the conjunctive case, $\kappa_{P_1}(C,E) > \kappa_{P_2}(C,E)$.
NO SUPERSESSION WITH DISJUNCTION: In the disjunctive case, $\kappa_{P_1}(C,E) = \kappa_{P_2}(C,E)$.

Again, we assume that probability in a Bayes net positively correlates not only with perceived statistical likelihood, but also with prescriptive normality. Thus, if event A is considered more normal in situation 1 than it is in situation 2, then we would expect $P_1(A) > P_2(A)$. So, under our interpretation of the probabilities as sampling propensities, these renderings capture the more informal descriptions of the effects

Necessity and Sufficiency

To determine *actual necessity* of $X = x$ for $Y = y$, we identify a path—a so called *active path*—from X to Y, and freeze all other variables $\vec{Z}$ outside the path to specific values $\vec{z}$. We then intervene to set $X \neq x$ and check whether nonetheless $Y = y$.

There are various proposals in the literature for how to choose $\vec{Z}$ and $\vec{z}$ (see, e.g., Halpern and Pearl 2005). For our purposes concerning the unshielded collider, all variables outside the (single) path will be fixed to their actual values. But in the general case, supposing we adopt some method for selecting $\vec{Z}$ and $\vec{z}$ in arbitrary models, we propose:

NECESSITY STRENGTH: $P\big(Y \neq y \mid do(X \neq x), \vec{Z} = \vec{z}\big)$

In the simple cases of interest here, the proposal amounts to setting C to 0, holding fixed $A = 1$, and checking whether this is enough to make $B = 0$; and likewise for assessing necessity strength of A. Thus, in the conjunctive model, the necessity strength of $C = 1$ is just $P\big(B = 0 \mid do(C = 0), A = 1\big) = 1$, and the necessity strength of $A = 1$ is also 1. In the disjunctive model, the necessity strengths are both 0 if in fact the other was present, while they are both 1 if the other was absent.

To determine *sufficiency* of $X = x$ for $Y = y$, we intervene to set $X = x$, forgetting everything about the actual situation, and sample forward to determine whether $Y = y$. That is:

SUFFICIENCY STRENGTH: $P\big(Y = y \mid do(X = x)\big)$

Again, in the unshielded collider model this value is quite simple. In the conjunctive case, the sufficiency strength of C is $P(A)$, for A it is $P(C)$. In the disjunctive case both are 1.

A Measure of Actual Causal Strength

Suppose X and Y are binary variables.[1] Then we can define a simple algorithm for determining causal strength of $X = 1$ on $Y = 1$ as in Figure 1. Intuitively, we simulate an X-situation,

1. Initialize $N = 0$, and for $k \leq K$:
   (a) Sample a value $X^{(k)}$ from $P(X)$.
   (b) If $X^{(k)} = 1$, draw $Y^{(k)}$ from $P\big(Y \mid do(X = 1)\big)$. Let $N = N + Y^{(k)}$.
   (c) If 0, draw $Y^{(k)}$ from $P\big(Y \mid do(X = 0), \vec{Z} = \vec{z}\big)$. Let $N = N + (1 - Y^{(k)})$.
2. Return N/K.

Figure 1: Algorithm for Determining Causal Strength

and depending on the value we sample for X, we either test for necessity or suffiency by simulating a Y-situation. It is easy to see that as $K \to \infty$, the fraction N/K converges to the following, which we take to be our measure of causal strength:

$$\begin{aligned}\kappa_P(X,Y) \quad = \quad & P(X = 1)\,P\big(Y = 1 \mid do(X = 1)\big) \; + \\ & P(X = 0)\,P\big(Y = 0 \mid do(X = 0), \vec{Z} = \vec{z}\big).\end{aligned}$$

The causal strength of $X = 1$ is simply the weighted sum of its sufficiency strength and necessity strength, these being weighted by $P(X = 1)$ and $P(X = 0)$, respectively.

It is possible for $\kappa_P(X,Y)$ to be arbitrarily close to 0, e.g., if the necessity strength of X is 0 but $P(X)$ is very small. But it cannot be 0 in any case where in fact $X = Y = 1$, since that would mean $P(X) > 0$ and that X clearly has some sufficiency strength. In general, $\kappa_P(X,Y) \in (0,1]$.

How does this account treat the three effects discussed earlier? Consider supersession. In the scenario with the coin and dice, the robust sufficiency of the coin coming up heads in the conjunctive case depends on how likely the dice are to sum to greater than the threshold. With a low threshold (2) this is quite likely, and hence robust sufficiency and overall causal strength are great; with a high threshold (11) it is unlikely, lessening the overall causal strength. Thus the statistical example goes through easily. The prescriptive case actually has exactly the same structure. Judgments of sufficiency strength for a given agent will depend on how often scenarios are sampled in which the other agent behaves differently. Under the assumption that this is more likely when the other agent violates a prescriptive norm we predict exactly the same pattern.

The other effects are explained similarly. In fact we have:

[1] The generalization is straightforward, if only cumbersome.

Fact 1. Given κ as a measure of causal strength, ABNORMAL SELECTION, SUPERSESSION, and NO SUPERSESSION WITH DISJUNCTION are all guaranteed.

While the precise numerical values will of course depend on the exact sampling propensities, Fact 1 shows that the patterns discussed above will be borne out so long as these probabilities reflect normality in the way that we have proposed.

It is worth emphasizing that no existing accounts of causal strength (e.g., among those surveyed by Fitelson and Hitchcock 2011) satisfy these desiderata. Not one of them can capture NO SUPERSESSION WITH DISJUNCTION. Pearl's (2009) measure of *Probability of Necessity and Sufficiency*, which is most similar to our measure κ, only captures SUPERSESSION. This is significant even if one is only interested in capturing people's causal judgments concerning purely probabilistic/statistical patterns, ignoring prescriptive normality altogether. The sampling view we propose naturally suggests a measure, κ, that easily captures all of these effects.

It is also worth mentioning that this account makes a striking further prediction about the simple 3-node causal structure we have been considering: namely, in disjunctive scenarios, the causal strength of a factor *C* should *decrease* with abnormality. Further work (together with Jonathan Kominksy) has confirmed this prediction in both the statistical and prescriptive cases. See Icard, Kominsky, and Knobe (2016) for details and further discussion of the present hypothesis.

Conclusion

We have offered a concrete model of actual causation judgments that accounts for the effect of normality—both statistical and prescriptive—by appeal only to elements and operations already present in the probabilistic graphical model representation of causal knowledge. While we believe the details of this particular algorithm and strength measure are of independent interest, our broader hypothesis, and central claim, is that the effect of normality on causal judgments can be explained by appeal to two main assumptions: (1) that in judging causal strength people probabilistically sample counterfactual states of the world, and (2) that these sampling probabilities are directly related to normality.

References

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

Fitelson, B. and Hitchcock, C. (2011). Probabilistic measures of causal strength. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 600–627. Oxford University Press.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Griffiths, T. L., Vul, E., and Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psych. Science*, 24(4):263–268.

Halpern, J. Y. and Hitchcock, C. (2015). Graded causation and defaults. *Brit. J. for Phil. Sci.*, 66(2):413–457.

Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. Part 1: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887.

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

Icard, T. F. (2016). Subjective probability as sampling propensity. *Rev. of Phil. and Psych.* forthcoming.

Icard, T. F., Kominsky, J. F., and Knobe, J. (2016). Causality, normality, and sampling propensity. Manuscript.

Kahneman, D. and Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psych. Rev.*, 94:136–153.

Knobe, J. and Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psych.*, 2:441–448.

Knobe, J. and Szabó, Z. (2008). Modals with a taste of the deontic. *Semantics and Pragmatics*, 6:1–42.

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

Lewis, D. (1973). Causation. *J. of Phil.*, pages 556–567.

Lieder, F., Hsu, M., and Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In *Proceedings of the 36th Annual Meeting in Cognitive Science*.

Livengood, J. and Rose, D. (2016). Experimental philosophy and causal attribution. In Sytsma, J. and Buckwalter, W., editors, *A Companion to Experimental Philosophy*. Blackwell. Forthcoming.

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

Lucas, C. G. and Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, 122(4):700–734.

McCloy, R. and Byrne, R. (2000). Counterfactual thinking and controllable events. *Memory and Cog.*, 28:1071–1078.

Pearl, J. (2009). *Causality*. Cambridge University Press.

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

Roxborough, C. and Cumby, J. (2009). Folk psychology concepts: Causation 1. *Philosophical Psych.*, 22(2):205–213.

Samland, J., Josephs, M., Waldmann, M. R., and Raboczy, H. (2016). The role of prescriptive norms and knowledge in childrens and adults causal selection. *Journal of Experimental Psychology: General*. forthcoming.

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