Word count (including references and notes): 7897

Experimental Philosophy and Causal Attribution

Jonathan Livengood
University of Illinois, Urbana-Champaign

David Rose
Rutgers University

Abstract. Humans often attribute the things that happen to one or another actual cause. In this chapter, we survey some recent philosophical and psychological research on causal attribution. We pay special attention to the relation between graphical causal modeling and theories of causal attribution. We think that the study of causal attribution is one place where formal and experimental techniques nicely complement one another.

Keywords: Causal attribution, attribution theory, actual causation, graphical causal model, experimental philosophy, default-deviant, norms, responsibility, blame

Acknowledgements: Thanks to Wesley Buckwalter, Tobi Gerstenberg, Joshua Knobe, Justin Sytsma, and two anonymous reviewers for helpful comments on earlier versions of this chapter.

Humans routinely solve a variety of different kinds of causal reasoning problems. In this chapter, we focus on problems having the following basic form. An agent has various bits of information about some things that happen in the world: the order in which things happen, the frequencies and conditional frequencies with which they happen, the things that are associated with interventions, and so on. The agent observes something happen, and she wants to know what caused that thing to happen. More specifically, she wants to know what *actually* caused the thing she observed to happen, not what *might* have caused it to happen or what *typically* causes similar things to happen. For example, she might want to know what actually caused her heartburn last night or she might want to know whether the bridge collapse was actually caused by microfractures in its box girders. Following Heider (1958), Jones and Davis (1965), Jones et al. (1972), Kelley (1967, 1971, 1972, 1973), Kelley and Michela (1980), and Weiner et al. (1971), we will call such problems *causal attribution problems*, though many other labels might have been appropriate as well: diagnostic inference problems, explanatory inference problems, and actual (or token) causation inference problems to name a few possible alternatives.

For Heider and the social psychologists influenced by him, attribution theory was an account of how people construct causal explanations, and the theory was primarily intended to describe how people explain the actions of others, e.g. by appeal to intentions, personality, situational factors, and so on.[i] Kelley (1973, 107) gives several examples of the kinds of questions of social perception that the theory was designed to handle, including the following (quoted verbatim):

> If a person is aggressively competitive in his behavior, is he this kind of person, or is he reacting to situational pressures?
>
> If a person advocates a certain political position, does this reflect his true opinions, or is it to be explained in some other way?

If a person fails on a test, does he have low ability, or is the test difficult? According to Kelley, "In all such instances, the *questions* concern the causes of observed behavior and the *answers* of interest are those given by the man in the street … what Heider has called 'naïve psychology.'" In the fifty years since Heider, psychologists and philosophers have made several suggestions about how ordinary causal cognition works. Various researchers have implicated ANOVA-like covariation (Kelley 1973), knowledge of mechanisms (Ahn et al. 1995), causal fields (Mackie 1965, 1974; Einhorn and Hogarth 1986), violations of normality (Hilton and Slugoski 1986; Knobe 2009; Hitchcock and Knobe 2009), and blameworthiness (Alicke 1992), to name just a few.

In this chapter, we survey several recent suggestions for understanding causal attribution, paying special attention to how the large body of research in attribution theory is related to recent work on graphical causal models. Here is how we will proceed. In Section 1, we situate causal attribution problems within a graphical causal modeling approach to causal reasoning. In Section 2, we review some recent research on structural approaches to causal attribution. In Section 3, we discuss a model that augments causal structure with a default-deviant distinction. In Sections 4 and 5 we discuss broadly normative considerations that influence causal attributions. Then we conclude in Section 6 with a discussion of some open questions and topics that we neglect in this chapter owing to the limitations of space.

1. Graphical Causal Models and Causal Attributions

Graphical causal models are an increasingly popular approach to thinking about causation in the philosophy and psychology literatures (see Burns and McCormack 2009; Danks THIS VOLUME; Fernbach and Sloman 2009; Glymour 2001; Glymour 2010; Gopnik et al. 2004;

Gopnik and Schulz 2007; Griffiths and Tenenbaum 2009; Lagnado and Sloman 2006; Lagnado et al. 2007; Park and Sloman 2013; Park and Sloman THIS VOLUME; Pearl 2000; Reips and Waldmann 2008; Rottman and Keil 2012; Rottman et al. 2014; Schulz et al. 2007; Sloman 2005; Sobel and Kushnir 2003; and Spirtes et al. 2000). In graphical causal modeling, we begin with a primitive relation of *direct structural causation* that takes variables as its relata. If a variable X is a direct structural cause of another variable Y with respect to some collection $\boldsymbol{V}$ of variables, then we write $X \rightarrow Y$ in a directed graph over $\boldsymbol{V}$. As an illustration, we will provide a graphical model for the following simple story. In a certain park, there are two clowns, Bozo and Zobo. Also, there are lots of peculiar children. What makes the children peculiar is the causal law that governs when they smile. Each child is such that she smiles just in case she receives a balloon from a clown. To model this story, we use three binary variables: *B*, *Z*, and *S*. For each child, the variable *B* represents whether or not Bozo gives the child a balloon, the variable *Z* represents whether or not Zobo gives the child a balloon, and the variable *S* represents whether or not the child smiles. According to the story, both *B* and *Z* are direct structural causes of *S*. And we represent that fact with the graph in Figure 1:

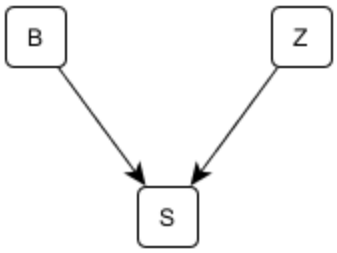

Figure 1: Causal Graph for the Clown Story

Formally, a *causal model* is a pair $<\boldsymbol{V}, \boldsymbol{F}>$ consisting of a collection $\boldsymbol{V}$ of variables and a collection $\boldsymbol{F}$ of functions relating elements of $\boldsymbol{V}$. For each variable $V \in \boldsymbol{V}$, there is a function $f_V \in \boldsymbol{F}$ that either specifies the value of V directly (in which case, V has no causes in the model and is

said to be *exogenous*) or specifies how the value of V is determined by the values of the direct structural causes of V. In the clown-smile story, we have the following equations:

$$Z = U_Z$$
$$B = U_B$$
$$S = B \vee Z$$

Where U_B and U_Z are the values that the functions f_B and f_Z assign to B and Z, respectively. (Typically, variables like U_B are interpreted as representing all of the unmeasured causes of the associated variable—in this case, B.)

We can give a non-reductive definition of direct structural causation in terms of ideal interventions that set the values of selected variables to specific values. The basic idea is to imagine testing whether one variable causes another by first holding every other variable fixed and then wiggling the first variable. If the second wiggles along, then the first causes the second. The formal construction looks like this. Suppose the variables X and Y are members of the collection $\boldsymbol{V}$ of all of the random variables in our model. Let $\boldsymbol{Z}$ be the collection $\boldsymbol{V} \setminus \{X, Y\}$, and let $Y_{\boldsymbol{W}=\boldsymbol{w}}$ denote the value Y would have if one were to set the variables in $\boldsymbol{W}$ to the values $\boldsymbol{w}$ by directly manipulating them. We can then say that X is a *direct structural cause* of Y relative to $\boldsymbol{V}$ iff there exists a vector $\boldsymbol{z}$ of values for the variables in $\boldsymbol{Z}$ and a pair of values x_1 and x_2 for X such that $Y_{\boldsymbol{Z}=\boldsymbol{z}, X=x1} \neq Y_{\boldsymbol{Z}=\boldsymbol{z}, X=x2}$. With these formal tools in hand, a causal attribution problem amounts to saying, for a pair of variables V_1 and V_2 evaluated with respect to some unit u, whether or not $V_1(u) = v_1$ is an actual cause of $V_2(u) = v_2$.

Consider Bozo and Zobo again. Suppose that Bozo, but not Zobo, gives little Suzy a balloon in the park, and Suzy smiles—as she must according to the structural equations:

$$Z(\text{Suzy}) = 0$$
$$B(\text{Suzy}) = 1$$
$$S(\text{Suzy}) = B \vee Z = 1$$

Where *B*(u) equals one if Bozo gives a balloon to unit u and equals zero if Bozo does not give a balloon to unit u. Where, similarly, *Z*(u) equals one of Zobo gives a balloon to unit u and equals zero if Zobo does not give a balloon to unit u. And where *S*(u) equals one if unit u smiles and equals zero if unit u does not smile. In this case, although both *B* and *Z* are structural causes of *S*, only *B*(Suzy) = 1 is an actual cause of *S*(Suzy) = 1.

Cases like the clown story are straightforward. Many will agree that Bozo—but not Zobo—actually caused Suzy to smile. However, as the story illustrates, the actual causes of a variable taking on the value it does are not always equivalent to the graphical ancestors of that variable or to the value(s) taken by those graphical ancestors. In order to identify the actual causes of a given variable taking on a specific value, we need more than just a list of the target variable's structural causes. But exactly what the *something more* should be turns out to be a very difficult question.

2. Actual Causation and Causal Structure

Several competing theories of actual causation have appealed to purely structural features of causal models as the extra something.[ii] The accounts have varying degrees of complexity, but the basic idea for each account is that a variable taking on some value is an actual cause of another variable taking on some value if there is some appropriate, possibly non-actual context in which the second variable taking on its actual value counterfactually depends on the first variable taking on its actual value. To illustrate how such proposals are supposed to work in a bit more detail, consider Woodward's (2003) account of actual causation.

When X is a direct structural cause of Y, we write $X \rightarrow Y$. A *path* of length $n > 0$ from a variable V_i to another variable V_j in a directed graph is a sequence $V_{(1)}, \ldots, V_{(n+1)}$ such that $V_i =$

$V_{(1)}$, $V_j = V_{(n+1)}$, and $V_{(k)} \rightarrow V_{(k+1)}$ for $k = 1, \ldots, n$. Let $\boldsymbol{W}$ denote an ordered n-tuple of variables, let $\boldsymbol{w}$ denote an ordered n-tuple of values of the variables in $\boldsymbol{W}$, and let the expression $do(\boldsymbol{W} = \boldsymbol{w})$ denote an ordered n-tuple of manipulations that set the variables in $\boldsymbol{W}$ to the values in $\boldsymbol{w}$. We say that $\boldsymbol{w}$ is in the *redundancy range* of the path P if carrying out the manipulations in $do(\boldsymbol{W} = \boldsymbol{w})$ leaves all of the variables on the path P at their actual values.

Now, according to Woodward (2003, 74-77), $X(\mathrm{u}) = x$ is an actual cause of $Y(\mathrm{u}) = y$ iff the following two conditions are satisfied:

(H1) The actual value of X is x and the actual value of Y is y, for unit u.

(H2) There exists a path P from X to Y and there exist manipulations $do(X = x^*)$ for $x^* \neq x$ and $do(\boldsymbol{W} = \boldsymbol{w})$ for $\boldsymbol{w}$ in the redundancy range of P such that $Y_{X=x^*}(\mathrm{u}) \neq y$ whenever the variables in $\boldsymbol{W}$ are fixed by the manipulation $do(\boldsymbol{W} = \boldsymbol{w})$.

In other words, $X(\mathrm{u}) = x$ is an actual cause of $Y(\mathrm{u}) = y$ if one can find some path P from X to Y and some choice of (possibly non-actual) values for all of the variables *not* on path P such that the variables on P retain their actual values and some change in the value of X would result in a change in the value of Y, if one were to set the variables not on path P to those values.

If we think of graphical modeling accounts of actual causation (like Woodward's) as models of naïve causal attributions, then they make predictions about what people will say in various cases. Though no one has published direct tests of these models, Livengood compared folk attributions of causation in a pilot study involving two simple voting scenarios. Each participant saw one of two vignettes describing a small election. In one vignette, every vote for the winning candidate is *pivotal* for the outcome, meaning that the result counterfactually depends on each of the votes for the winning candidate. In the other vignette, the outcome is over-determined: the result does not counterfactually depend on any single vote. The vignette with counterfactual dependence reads like this:

Thirteen votes were cast in an election involving three candidates, Smith, Jones, and Murphy. The vote totals were as follows:

Smith	6
Jones	5
Murphy	2

Greg voted for Smith. Was Greg's vote a cause of Smith winning the election? [yes / no]

The other vignette was identical except that the 13 votes were assigned differently: ten for Smith, two for Jones, and one for Murphy. The relative percentage of "yes" answers for each of the two vignettes is pictured in Figure 2.

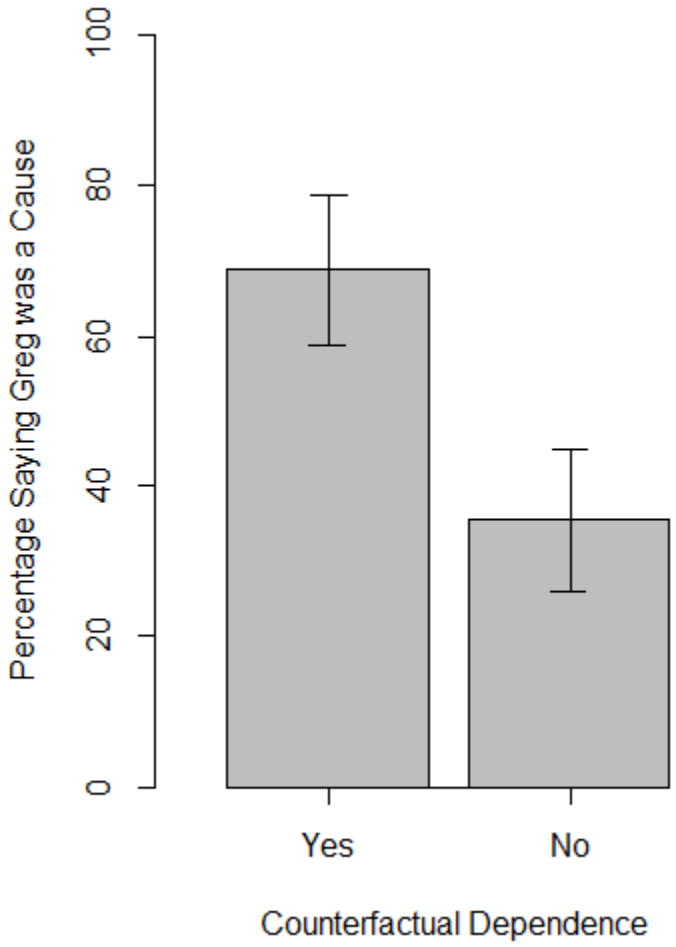

Figure 2: Causal Attributions with and without Counterfactual Dependence

Understood as models of ordinary causal attributions, all of the graphical accounts predict that Greg's vote will be counted as a cause of Smith winning the election *regardless of whether the outcome counterfactually depends on Greg's vote*.[iii] But people do not treat Greg's vote the same way in both cases.[iv] Livengood's study raises some doubt about the adequacy of graphical models of actual causation as accounts of naïve causal attribution, but it is hardly definitive.

A different graphical modeling proposal by Chockler and Halpern (2004) has been more extensively tested. The main idea is to measure the degree of causal responsibility of a given variable's value in terms of the number of changes that would need to be made to the actual model in order to make the target variable's value pivotal. In the counterfactual-dependence condition of the voting experiment, we do not need to make any changes to the model for Greg's vote to be pivotal. Greg's vote is already pivotal in the actual model. By contrast, in the no-dependence condition we need to make three changes—move three votes from Smith to Jones—in order to make Greg's vote pivotal.

Chockler and Halpern (2004) propose to measure the degree of causal responsibility of $X(u) = x$ for $Y(u) = y$ according to the equation

$$\deg(X = x, Y = y) = \frac{1}{N+1}$$

where deg is degree of causal responsibility, and N is the minimal number of changes needed in order to make $X(u) = x$ pivotal with respect to $Y(u) = y$. In Livengood's election vignettes, Greg's degree of causal responsibility is 1 in the dependence condition and ¼ in the no-dependence condition.

Gerstenberg and Lagnado (2010) tested Chockler and Halpern's proposal against two other models—the *counterfactual* model and the *matching* model—with *the Triangle Game*. In the Triangle Game, participants are given a short period of time to count the number of triangles

in a complex display. Participants played as part of a group, and the conditions under which a team won or lost were manipulated. For example, winning might require *all* of the players to give answers close enough to the truth, or winning might require *at least one* of the players to give an answer close enough to the truth. After answering, participants saw the correct answer and the answers given by the other players. Then they were asked to rate each player's degree of responsibility for the team's win or loss.

Gerstenberg and Lagnado looked at how well the predictions of the three models correlated with the responsibility ratings of their participants. They found that the median correlation between model and participant was greatest for the structural model and that the structural model had the best fit to the ratings of 52 of their 69 participants. Lagnado et al. (2013) improve on the model by incorporating a second structural feature: how important some variable's value is expected to be before any of the values are known.

The structural models considered by Lagnado and colleagues are designed to handle attributions of causal responsibility when many variables contributed to some outcome, which limits their applicability. A more serious limitation, which plagues structural models of actual causation generally, is the threat of *isomorphisms* (Hall 2007; Halpern and Hitchcock forthcoming). To illustrate, consider the following two cases due to Hiddleston (2005):

> *Overdetermination*: Billy and Suzy both throw a rock at a window at the same time. Both rocks reach the window, shattering it upon impact.
>
> *Bogus Prevention*: Killer plans to poison Victim's coffee, but has a change of heart and refrains from administering the lethal poison. Bodyguard puts an antidote in the coffee that would have neutralized the poison (had there been any present). Victim drinks the coffee and (of course) survives.

Simple graphical models of both Overdetermination and Bogus Prevention have the same v-shaped causal structure as the model of Bozo and Zobo in Section 1. They're structurally

isomorphic. Hence, any purely structural account of actual causation must treat Billy, Suzy, Bozo, Zobo, Killer, and Bodyguard exactly alike. Yet, to many it has seemed that in Overdetermination, both Billy and Suzy are actual causes of the window shattering, while in Bogus prevention, Bodyguard is *not* an actual cause of Victim surviving.[v]

3. Modeling the Default-Deviant Distinction

Many researchers (e.g. Menzies 2004, 2007; Hall 2007; Halpern and Hitchcock forthcoming; Hitchcock 2007; Hitchcock and Knobe 2009; Livengood 2013) have inferred from the problem of isomorphisms and other considerations that purely structural accounts of actual causation need to be supplemented with a default-deviant distinction. Identifying some values as default and some as deviant would allow modelers to distinguish isomorphic causal models and better capture ordinary causal attributions. But there are many different ways to augment structural models with a default-deviant distinction. In this section, we describe Hitchcock's (2007) attempt to incorporate defaults into graphical causal models.

Let $<\boldsymbol{V}, \boldsymbol{F}>$ be a causal model, and let $X, Y \in \boldsymbol{V}$. Define a *causal network* connecting X to Y in $<\boldsymbol{V}, \boldsymbol{F}>$ to be the set $\boldsymbol{N} \subseteq \boldsymbol{V}$ that contains exactly X, Y and all variables Z in $\boldsymbol{V}$ lying on a path from X to Y in $<\boldsymbol{V}, \boldsymbol{F}>$. Say that a causal network $\boldsymbol{N}$ connecting X to Y is *self-contained* iff for all $Z \in \boldsymbol{N}$, if Z has parents in $\boldsymbol{N}$, then Z takes a default value when all of its parents in $\boldsymbol{N}$ take their default values and all of its parents in $\boldsymbol{V} \setminus \boldsymbol{N}$ take their actual values. According to Hitchcock, counterfactual dependence is necessary and sufficient for actual causation in a self-contained network, a claim he formalizes as follows:

> TC: Let $<\boldsymbol{V}, \boldsymbol{F}>$ be a causal model, let $X, Y \in \boldsymbol{V}$, and let $X = x$ and $Y = y$. If the causal network connecting X to Y in $<\boldsymbol{V}, \boldsymbol{F}>$ is self-contained, then $X = x$ is an actual cause of $Y = y$ in $<\boldsymbol{V}, \boldsymbol{F}>$ if and only if the value of Y counterfactually depends on the value of X in $<\boldsymbol{V}, \boldsymbol{F}>$.

If TC is a correct description of the psychology of causal attribution, we can make predictions provided we have the right causal model and the right choice of default values for the variables in the model.[vi]

Livengood et al. (ms) tested Hitchcock's TC using modified versions of a thought experiment due to Knobe. In one experiment, participants read a story about Lauren and Jane, who work at a company with an unstable computer system such that if more than one person logs in at the same time, the system crashes. Participants are told that one day, both women log into the system at the same time, and it crashes. They were then asked to rate their level of agreement with the following three claims: (1) Lauren caused the system to crash, (2) Jane caused the system to crash, and (3) the instability in the system caused it to crash. The results are pictured in Figure 3.

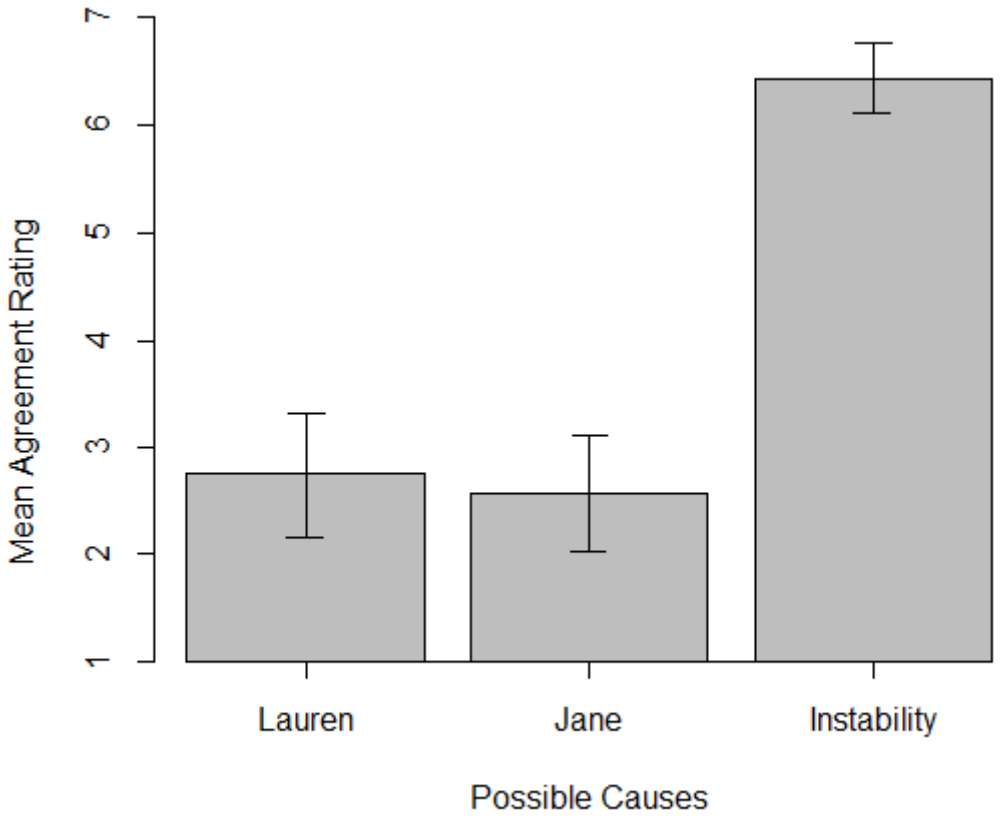

Figure 3: Causal Attributions in Lauren and Jane Experiment

Whether the data confirm or disconfirm TC as a model of causal attribution depends crucially on the choice of default values for the variables. If the default values for Lauren and Jane are "logs in," then the experiment confirms Hitchcock's account. But if the default values are "does not log in," then the experiment is disconfirming. Livengood et al. argue that on Hitchcock's account, the default value for Lauren and for Jane is "does not log in," since the act of logging in is a voluntary departure from a rest state. But plausibly, participants regarded the actions of Lauren and Jane as having default status because they were in some sense normal actions in the circumstances.[vii] Such a possibility calls for more research on how people make normality judgments and on which kinds of normality judgments matter for causal attributions. We take up the latter topic in the next section.

4. Varieties of Norms and Their Influence

Hitchcock and Knobe (2009) suggest that normative considerations matter for causal attributions in virtue of the fact that paying attention to what is abnormal helps an agent to choose which counterfactuals to evaluate in a given context. The distinction between causes and mere background conditions, for example, depends on judgments of normality. Normal states of affairs are regarded as potential enabling background conditions; whereas, abnormal states of affairs are regarded as potential causes. Hitchcock and Knobe were not the first to draw a connection between causation and abnormality: Hart and Honoré (1959) discuss the role of abnormality in causal attribution in the law; Hilton and Slugoski (1986) discuss the interplay between norm-violation and information called on by covariational models of causal attribution; and Kahneman and Miller (1986) discuss the role of category norms in causal attributions.[viii] The main novelty in Hitchcock and Knobe's theory—as we understand it—is that they explicitly treat norms and norm-violations as including much more than statistical facts or facts about how well one exemplifies membership in a natural kind or category.

Hitchcock and Knobe argue that only overall judgments of normality matter for causal attributions. However, they distinguish three types of norms relevant to causal attributions: statistical norms, prescriptive norms, and norms of proper functioning. Statistical norms have to do with what is typical or atypical. For example, a lightning strike in a forest is atypical, violating a statistical norm. But the presence of oxygen in the forest is typical, conforming to a statistical norm. By contrast, prescriptive norms have to do with what is *right* or *wrong*. For example, jaywalking violates a prescriptive norm, even if people regularly do so. And telling the complete truth to the police conforms to a prescriptive norm, even if people only rarely do so.

Finally, norms of proper functioning concern the behavior of mechanisms designed or selected to do a specific thing. For example, a smoke detector that beeps just in case there is smoke conforms to a norm of proper functioning, while a bicycle with a stuck brake violates a norm of proper functioning. In what follows, we will consider some recent evidence regarding the extent of the influence of various normative considerations on causal attributions.

Evidence that normative considerations affect judgments of actual causation comes from a range of studies, although the studies do not always support the theoretical picture advocated by Hitchcock and Knobe. Knobe and Fraser (2009) provided evidence that ordinary causal attributions are influenced by prescriptive norms in their *pen 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 repeatedly e-mails 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.

When asked to indicate the extent to which the administrative assistant and the professor caused the problem, participants were much more likely to indicate that the professor was a cause of the problem. Further evidence for the claim that prescriptive norms influence causal attributions is adduced in Kominsky et al. (2014).

Hitchcock and Knobe (2009) provide some evidence that norms of proper functioning also matter to causal attribution in their *wires case*:

> A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is

> supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine.
>
> One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit. (p. 604)

After reading the wires case, participants were asked to indicate the extent to which they thought the red or black wires touching the battery caused the machine to short circuit. Hitchcock and Knobe found that people were more willing to say that the red wire's touching the battery was an actual cause of the short circuit.

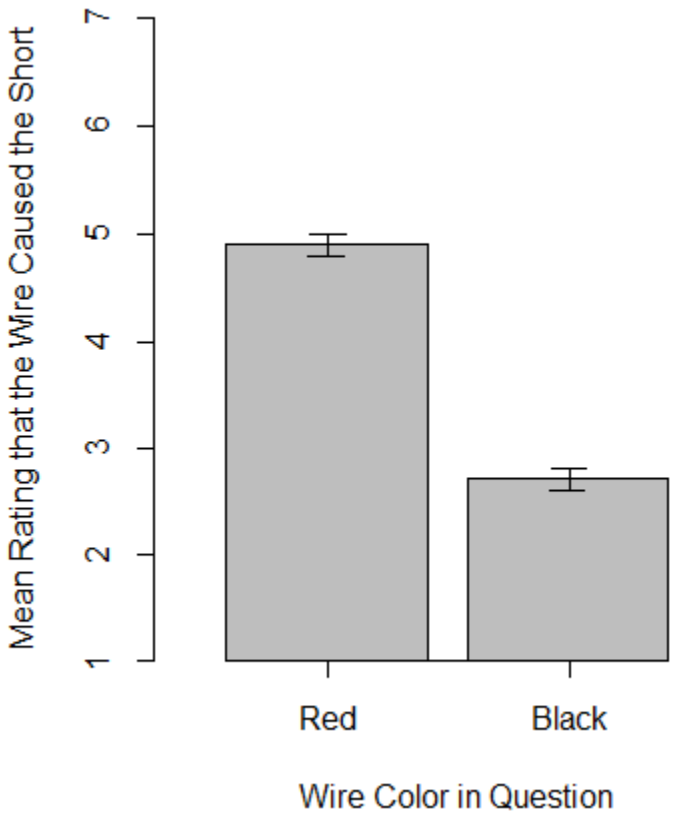

Figure 4: Hitchcock and Knobe (2009) Wires Case

Hitchcock and Knobe explain their finding by appealing to the fact that according to its design, the red wire is not supposed to be touching the battery.

Roxborough and Cumby (2009) modified Knobe and Fraser's pen case so that the administrative assistants do *not* typically take pens. They found that participants in their study made lower causal ratings for the professor than did participants in Knobe and Fraser's study. Roxborough and Cumby take this to provide support for the claim that statistical norm violations do, indeed, affect folk judgments of actual causation. Sytsma et al. (2012) distinguish two sub-types of statistical norm: population-level and agent-level.[ix] Sytsma et al. found that causal attributions were not sensitive to violations of population-level statistical norms and that while the agent-level statistical norms they tested mattered, they did so in the exact opposite way as that predicted by Hitchcock and Knobe: agent-typical behaviors were more likely to be judged as causes than were agent-atypical behaviors. The upshot is that we have good evidence that normative considerations affect causal attributions, but we do not currently have a good theoretical model describing precisely how they do so.

5. Causal Attributions and the Desire to Blame

A further wrinkle in providing an adequate account of causal attribution is the influence of a desire to blame on causal attributions. The best current account of the way praise and blame figure in causal attribution is Alicke's Culpable Control Model (Alicke 1992, 2000; Alicke and Rose 2010; Alicke, Rose, and Bloom 2011). According to the CCM, in the realm of harmful and offensive actions, ordinary causal attributions are biased by a desire to blame those who we evaluate negatively. We exaggerate an actor's causal role in bringing about an event since doing so allows us to support our desire to blame the actor.[x]

In support of the CCM, Alicke, Rose, and Bloom (2011) conducted experiments suggesting that blame judgments *cause* ordinary causal attributions. Participants read a story in

which a character named Edward Poole is shot by a character named Mr. Turnbull in Turnbull's home. Alicke et al. varied whether Poole was characterized positively or negatively and the mode of Poole's death. Participants in the positive condition were told that Poole was a physician who was house-sitting for the Turnbulls while they were out of town. Participants in the negative condition were told that Poole was an ex-convict who had broken into the house.

In each of the positive and negative characterization conditions, participants were told that Mr. Turnbull shot Poole in the chest. Each participant was told one of three things about the gunshot and Poole's death—that the shot killed Poole instantly or that the shot killed Poole but he had an inoperable terminal brain tumor or that Poole had an aneurysm at almost the same time that he was shot. In all conditions, participants rated the extent to which Mr. Turnbull was the cause of Poole's death and the extent to which Mr. Turnbull was deserving of blame. Alicke et al. found that ratings of blame statistically screen off the way Poole is characterized (positively or negatively) from causal ratings, indicating that blame ratings mediate the effect of Poole's characterization on causal attributions. Moreover, they found that although the mode of Poole's death was independent of the way Poole was characterized, those two variables were dependent conditional on causal attributions. They thus inferred that the correct causal model for participants in their experiment is as pictured in Figure 5.

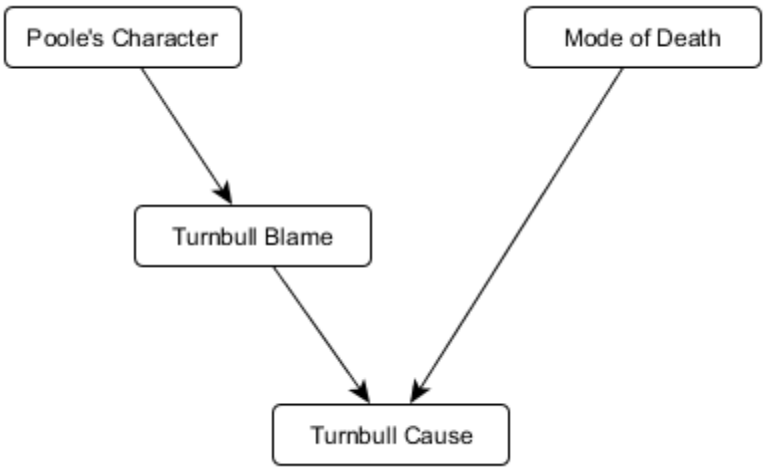

Figure 5: Causal Graph for Poole Experiment

Since the pattern of counterfactual dependence is the same in both the positive and negative conditions, the Poole Experiment suggests that in some cases, what matters for causal attribution is not the salience of various counterfactuals—as maintained by the structural and norm-violation views we have seen so far—but the desire to blame. Experiments like the Poole Experiment raise a difficult problem in many cases as to whether the observed causal attributions are due to sensitivity to norms or rather are due to sensitivity to a desire to blame. Hence, in many instances, the CCM is in competition with norm-violation accounts of causal attribution, though we think that it is possible for some version of both views to be correct.

6. Open Questions and Neglected Topics

We are now nearly out of space, so it's time to wrap things up. We have seen that graphical models offer an interesting way to unify much ongoing research on the problem of causal attribution. We considered some purely structural models of causal attribution and some of their limitations. We looked at one attempt to augment structural models with a default-deviant distinction and one serious modeling challenge for such approaches. We reviewed some research

suggesting that norms influence causal attributions and some research suggesting that causal attributions are biased by a desire to blame people. The questions addressed by the research we have reviewed are mostly still open. In closing, we want to mention a few more open questions and some issues that we did not have time to talk about in any detail.

One big question is the scope of the influence of norms on causal attribution. Causal attributions have been shown to happen in a wide range of cases that seem to lack norms of the sort that figure in Hitchcock and Knobe's account. These range from cases of launching and causal perception (Michotte 1963; Scholl and Tremoulet 2000; Scholl and Nakayama 2002), to related force dynamics cases (Talmy 1988; Wolff 2007), to attributions based on touch (Wolff et al. ms), to covariational cases (Danks et al. 2013), to simple vignette cases like Livengood's voting experiment. Normative considerations matter, but they are not the whole story for causal attribution. And it is an open question exactly when and why they matter.

Another big question hinted at but not explicitly raised in this chapter is whether sensitivity of causal attributions to normative considerations, the desire to blame or praise, and so on is best understood as a standard to be embraced or as a bias to be avoided. Two related issues come up at this point. The first is whether the concept of causation itself has normative content. If the ordinary concept of causation is fundamentally normative in character (as suggested by McGrath 2005 and in a different way by Sytsma et al. 2012), then it is unsurprising and (perhaps) unthreatening that ordinary causal attributions are sensitive to normative considerations. But now the second issue arises. What exactly is the point of attributing causation? By what standard should we judge the success or failure of causal attributions? What benefit accrues to an agent in virtue of her ability to solve a causal attribution problem? Some attention has been paid to these and related questions, but much more research needs to be done

(see Hitchcock and Knobe 2009 and Danks 2013 for illuminating work on these issues and Fisher 2014 for a basic framework within which experimental philosophy of this sort might proceed). Related to these concerns is another issue that arises in many areas of experimental philosophy: namely, to what extent should we trust ordinary intuitions (whatever those are) about causation? The answer may very well depend on an interaction between the shape our concept takes and the ends to which we put that concept (see Korman 2009; Rose 2015; and Rose and Schaffer ms, for discussions of these issues in a different setting).

Owing to limitations of space, we have not been able to say anything about experimental work on causation by absence or omission (Livengood and Machery 2007; Wolff et al. 2010; and references therein), causal explanation (Livengood and Machery 2007; Lombrozo 2006, 2007, 2010; Lombrozo and Carey 2006); the significance of causal language (Talmy 1988; Wolff et al. 2005; Wolff and Song 2003); or the relationship between causal attribution and judgments of moral or legal responsibility (Gerstenberg and Lagnado 2010, 2012) to name just a few of many topics related to causal attribution.

References

Ahn, W., C. Kalish, D. Medin, and S. Gelman. 1995. "The Role of Covariation Versus Mechanism Information in Causal Attribution." *Cognition*, 54: 299-352.
DOI: 10.1016/0010-0277(94)00640-7

Alicke, M., and D. Rose. 2010. "Culpable Control or Moral Concepts?" *Behavioral and Brain Sciences*, 33: 330-331.
DOI: 10.1017/S0140525X10001664

Alicke, M., D. Rose, and D. Bloom. 2011. "Causation, Norm Violation, and Culpable Control." *Journal of Philosophy*, 108: 670-696.

Blanchard, T., and J. Schaffer. Forthcoming. "Cause without default." In *Making a Difference*, edited by H. Beebee, C. Hitchcock, and H. Price. Oxford: Oxford University Press.

Burns, P., and T. McCormack. 2009. "Temporal Information and Children's and Adults' Causal Inferences." *Thinking & Reasoning*, 15: 167-196.
DOI: 10.1080/13546780902743609

Chockler, H., and J. Halpern. 2004. "Responsibility and Blame: A Structural-Model Approach." *Journal of Artificial Intelligence Research*, 22: 93-115.

Collins, J., N. Hall, and L. Paul, eds. 2004. *Causation and Counterfactuals*. Cambridge: MIT Press.

Danks, D. 2009. "The Psychology of Causal Perception and Reasoning." In *The Oxford Handbook of Causation*, edited by H. Beebee, C. Hitchcock, and P. Menzies, 447-470. Oxford: Oxford University Press.

Danks, D. 2013. "Functions and Cognitive Bases for the Concept of Actual Causation." *Erkenntnis*, 78: 111-128.
DOI: 10.1007/s10670-013-9439-2

Danks, D. 2015. "Causal Search, Causal Modeling, and the Folk." In THIS VOLUME.

Danks, D., D. Rose, and E. Machery. 2014. "Demoralizing Causation." *Philosophical Studies*, 171: 251-277.
DOI: 10.1007/s11098-013-0266-8

Einhorn, H., and R. Hogarth. 1986. "Judging Probable Cause." *Psychological Bulletin*, 99: 3-19.
DOI: 10.1037/0033-2909.99.1.3

Fernbach, P., and S. Sloman. 2009. "Causal Learning with Local Computations." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35: 678-693.
DOI: 10.1037/a0014928

Fisher, J. 2014. “Pragmatic Experimental Philosophy.” *Philosophical Psychology*, X: 1-22. DOI: 10.1080/09515089.2013.870546

Gerstenberg, T., and D. Lagnado. 2010. “Spreading the Blame: The Allocation of Responsibility Amongst Multiple Agents.” *Cognition*, 115: 166-171. DOI: 10.1016/j.cognition.2009.12.011

Gerstenberg, T., and D. Lagnado. 2012. “When Contributions Make a Difference: Explaining Order Effects in Responsibility Attribution.” *Psychonomic Bulletin & Review*, 19: 729-736. DOI: 10.3758/s13423-012-0256-4

Glymour, C. 2001. *The Mind’s Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge: MIT Press.

Glymour, C. 2010. “What Is Right with ‘Bayes Net Methods’ and What Is Wrong with ‘Hunting Causes and Using Them’?” *The British Journal for the Philosophy of Science*, 61: 161-211. DOI: 10.1093/bjps/axp039

Glymour, C., D. Danks, B. Glymour, F. Eberhardt, J. Ramsey, R. Scheines, and J. Zhang. 2010. “Actual Causation: A Stone Soup Essay.” *Synthese*, 175: 169-192. DOI: 10.1007/s11229-009-9497-9

Gopnik, A., C. Glymour, D. Sobel, L. Schulz, T. Kushnir, and D. Danks. 2004. “A Theory of Causal Learning in Children: Causal Maps and Bayes Nets.” *Psychological Review*, 111: 3-32. DOI: 10.1037/0033-295X.111.1.3

Gopnik, A., and L. Schulz, eds. 2007. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780195176803.001.0001

Griffiths, T., and J. Tenenbaum. 2009. “Theory-Based Causal Induction.” *Psychological Review*, 116: 661. DOI: 10.1037/a0017201

Hall, N. 2007. “Structural Equations and Causation.” *Philosophical Studies*, 132: 109-136. DOI: 10.1007/s11098-006-9057-9

Halpern, J. 2008. “Defaults and Normality in Causal Structures.” In *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference*, 198-208.

Halpern, J., and C. Hitchcock. Forthcoming. “Graded Causation and Defaults.” *The British Journal for the Philosophy of Science*. DOI:10.1093/bjps/axt050

Halpern, J., and J. Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part I: Causes." *The British Journal for the Philosophy of Science*, 56: 843-887.
DOI: 10.1093/bjps/axi147

Hart, H. and T. Honoré. 1959. *Causation in the Law*. Oxford: Clarendon Press.
DOI: 10.1093/acprof:oso/9780198254744.001.0001

Heider, F. 1958. *The Psychology of Interpersonal Relations*. New York: Wiley.
DOI: 10.1037/10628-000

Hiddleston, E. 2005. "A Causal Theory of Counterfactuals." *Noûs*, 39: 632-657.
DOI: 10.1111/j.0029-4624.2005.00542.x

Hilton, D., and B. Slugoski. 1986. "Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model." *Psychological Review*, 93: 75-88.
DOI: 10.1037/0033-295X.93.1.75

Hitchcock, C. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy*, 98: 273-299.
DOI: 10.2307/2678432

Hitchcock, C. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review*, 116: 495-531.
DOI: 10.1215/00318108-2007-012

Hitchcock, C., and J. Knobe. 2009. "Cause and Norm." *Journal of Philosophy*, 106: 587-612.

Jones, E., and K. Davis. 1965. "From Acts to Dispositions: The Attribution Process in Person Perception." *Advances in Experimental Social Psychology*, 2: 219-266.
DOI: 10.1016/S0065-2601(08)60107-0

Kahneman, D., and D. Miller. 1986. "Norm Theory: Comparing Reality to Its Alternatives." *Psychological Review*, 80: 136-153.
DOI: 10.1037/0033-295X.93.2.136

Kanouse, D., H. Kelley, R. Nisbett, S. Valins, B. Weiner, and E. Jones. 1972. *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Kelley, H. 1967. "Attribution Theory in Social Psychology." In *Nebraska Symposium on Motivation*. University of Nebraska Press.

Kelley, H. 1971. *Attribution in Social Interaction*. New York: General Learning Press.

Kelley, H. 1972. *Causal Schemata and the Attribution Process*. Morristown, NJ: General Learning Press.

Kelley, H. 1973. “The Processes of Causal Attribution.” *American Psychologist*, 28: 107.
DOI: 10.1037/h0034225

Kelley, H., and J. Michela. 1980. “Attribution Theory and Research.” *Annual Review of Psychology*, 31: 457-501.
DOI: 10.1146/annurev.ps.31.020180.002325

Knobe, J. 2009. “Folk Judgments of Causation.” *Studies in History and Philosophy of Science Part A*, 40: 238-242.
DOI: 10.1016/j.shpsa.2009.03.009

Knobe, J. 2010. “Person as Scientist, Person as Moralist.” *Behavioral and Brain Sciences,* 33: 315-329.
DOI: 10.1017/S0140525X10000907

Knobe, J., and B. Fraser. 2008. “Causal Judgment and Moral Judgment: Two Experiments.” In *Moral Psychology: Volume 2*, edited by W. Sinnott-Armstrong, 441-448. Cambridge: MIT Press.

Kominsky, J., J. Phillips, T. Gerstenberg, D. Lagnado, and J. Knobe. MS. “Causal Supersession.” http://web.mit.edu/tger/www/papers/Causal%20supersession%20(Kominsky%20et%20al,%202014).pdf

Korman, D. 2009. “Eliminativism and the Challenge from Folk Belief.” *Noûs*, 43: 242-264.
DOI: 10.1111/j.1468-0068.2009.00705.x

Lagnado, D., and S. Sloman. 2006. “Time as a Guide to Cause.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32: 451.
DOI: 10.1037/0278-7393.32.3.451

Lagnado, D., M. Waldmann, Y. Hagmayer, and S. Sloman. 2007. “Beyond Covariation.” In *Causal Learning: Psychology, Philosophy, and Computation*, edited by A. Gopnik and L. Schulz, 154-172. Oxford: Oxford University Press.
DOI: 10.1093/acprof:oso/9780195176803.003.0011

Lagnado, D., T. Gerstenberg, and R. Zultan. 2013. “Causal Responsibility and Counterfactuals.” *Cognitive Science*, 37: 1036-1073.
DOI: 10.1111/cogs.12054

Livengood, J., and E. Machery. 2007. “The Folk Probably Don't Think What You Think They Think: Experiments on Causation by Absence.” *Midwest Studies in Philosophy*, 31: 107–127.
DOI: 10.1111/j.1475-4975.2007.00150.x

Livengood, J. 2013. “Actual Causation and Simple Voting Scenarios.” *Noûs* 47: 316-345.
DOI: 10.1111/j.1468-0068.2011.00834.x

Livengood, J., J. Sytsma, and D. Rose. MS. “Following the FAD: Folk Attributions and Theories of Actual Causation.” http://jonathanlivengood.net/Folk%20Attributions%20and%20Theories%20of%20Actual%20Causation__ms.pdf

Lombrozo, T. 2006. “The Structure and Function of Explanations.” *Trends in Cognitive Sciences*, 10: 464-470. DOI: 10.1016/j.tics.2006.08.004

Lombrozo, T. 2007. “Simplicity and Probability in Causal Explanation.” *Cognitive Psychology*, 55: 232-257. DOI: 10.1016/j.cogpsych.2006.09.006

Lombrozo, T. 2010. “Causal–Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions.” *Cognitive Psychology*, 61: 303-332. DOI: 10.1016/j.cogpsych.2010.05.002

Lombrozo, T., and S. Carey. 2006. “Functional Explanation and the Function of Explanation.” *Cognition*, 99: 167-204. DOI: 10.1016/j.cognition.2004.12.009

Mackie, J. 1965. “Causes and Conditions.” *American Philosophical Quarterly*, 2: 245-264.

Mackie, J. 1974. *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon Press.

Malle, B. 2011. “Time to Give Up the Dogmas of Attribution: An Alternative Theory of Behavior Explanation.” *Advances in Experimental Social Psychology*, 44: 297-311. DOI: 10.1016/B978-0-12-385522-0.00006-8

McGrath, S. 2005. “Causation by Omission: A Dilemma.” *Philosophical Studies*, 123: 125-148. DOI: 10.1007/s11098-004-5216-z

Menzies, P. 2004. “Difference-Making in Context.” In *Causation and Counterfactuals*, edited by J. Collins, N. Hall, and L. Paul, 139-180. Cambridge: MIT Press.

Menzies, P. 2007. “Causation in Context.” In *Causation, Physics, and the Constitution of Reality: Russell’s Republic Revisited*, edited by H. Price and R. Cory, 191-223. Oxford: Oxford University Press.

Park, J., and S. Sloman. 2013. “Mechanistic Beliefs Determine Adherence to the Markov Property in Causal Reasoning.” *Cognitive Psychology*, 67: 186-216. DOI: 10.1016/j.cogpsych.2013.09.002

Park, J. and S. Sloman. 2015. “Causal Models and Screening-Off.” In THIS VOLUME

Pearl, J. 2000. *Causality: Models, Reasoning and Inference*. Cambridge: MIT Press.

Reips, U., and M. Waldmann. 2008. "When Learning Order Affects Sensitivity to Base Rates: Challenges for Theories of Causal Learning." *Experimental Psychology*, 55: 9-22.
DOI: 10.1027/1618-3169.55.1.9

Rose, D. 2015. "Persistence through Function Preservation." *Synthese*, 192: 97-146.
DOI: 10.1007/s11229-014-0555-6

Rose, D., and D. Danks. 2012. "Causation: Empirical Trends and Future Directions." *Philosophy Compass*, 7: 643-653.
DOI: 10.1111/j.1747-9991.2012.00503.x

Rose, D., and J. Schaffer. MS. "Folk Mereology is Teleological."

Rottman, B., and F. Keil. 2012. "Causal Structure Learning over Time: Observations and Interventions." *Cognitive Psychology*, 64: 93-125.
DOI: 10.1016/j.cogpsych.2011.10.003

Rottman, B., J. Kominsky, and F. Keil. 2014. "Children Use Temporal Cues to Learn Causal Directionality." *Cognitive Science*, 38: 489-513.
DOI: 10.1111/cogs.12070

Roxborough, C., and J. Cumby. 2009. "Folk Psychological Concepts: Causation." *Philosophical Psychology*, 22: 205-213.
DOI: 10.1080/09515080902802769

Scholl, B., and K. Nakayama. 2002. "Causal Capture: Contextual Effects on the Perception of Collision Events." *Psychological Science*, 13: 493-498.
DOI: 10.1111/1467-9280.00487

Scholl, B., and P. Tremoulet. 2000. "Perceptual Causality and Animacy." *Trends in Cognitive Sciences*, 4: 299-309.
DOI: 10.1016/S1364-6613(00)01506-0

Schulz, L., T. Kushnir, and A. Gopnik. 2007. "Learning from Doing: Intervention and Causal Inference." In *Causal Learning: Psychology, Philosophy, and Computation*, edited by A. Gopnik and L. Schulz, 67-85. Oxford: Oxford University Press.
DOI: 10.1093/acprof:oso/9780195176803.003.0006

Sloman, S. 2005. *Causal Models: How People Think about the World and its Alternatives*. Oxford: Oxford University Press.
DOI: 10.1093/acprof:oso/9780195183115.001.0001

Sobel, D., and T. Kushnir. 2003. "Interventions Do Not Solely Benefit Causal Learning: Being Told What to do Results in Worse Learning than Doing it Yourself." In *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*.

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search*. Cambridge: MIT Press.

Sytsma, J., J. Livengood, and D. Rose. 2012. "Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions." *Studies in History and Philosophy of Science Part C*, 43: 814-820.

Talmy, L. 1988. "Force Dynamics in Language and Cognition." *Cognitive Science*, 12: 49-100. DOI: 10.1207/s15516709cog1201_2

Weiner, B. 1971. *Perceiving the Causes of Success and Failure*. New York: General Learning Press.

Wolff, P. 2007. "Representing Causation." *Journal of Experimental Psychology: General*, 136: 82-111.
DOI: 10.1037/0096-3445.136.1.82

Wolff, P., A. Barbey, and M. Hausknecht. 2010. "For Want of a Nail: How Absences Cause Events." *Journal of Experimental Psychology: General*, 139: 191-221.
DOI: 10.1037/a0018129

Wolff, P., B. Klettke, T. Ventura, and G. Song. 2005. "Expressing Causation in English and Other Languages." In *Categorization Inside and Outside the Laboratory: Essays in Honor of Douglas L. Medin*, edited by W. Ahn, R. Goldstone, B. Love, A. Markman, and P. Wolff, 29-48. Washington, D.C.: American Psychological Association.
DOI: 10.1037/11156-003

Wolff, P., S. Ritter, and K. Holmes. MS. "Causation, Force, and the Sense of Touch." https://mindmodeling.org/cogsci2014/papers/310/paper310.pdf

Wolff, P., and G. Song. 2003. "Models of Causation and the Semantics of Causal Verbs." *Cognitive Psychology*, 47: 276-332.
DOI: 10.1016/S0010-0285(03)00036-7

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Endnotes

[i] See Malle (2011) for an argument that psychologists have mostly misread Heider.

[ii] See Hitchcock (2001); Woodward (2003); and Halpern and Pearl (2005) for examples. See Glymour et al. (2010) and Livengood (2013) for critical discussion.

[iii] The graphical models say (roughly) that since there is a way of redistributing the votes such that under the redistribution, the outcome counterfactually depends on Greg's vote, Greg's vote is an actual cause of the outcome. See Livengood (2013) for much more detail.

[iv] A total of 196 participants were recruited on the Philosophical Personality website and randomly assigned to either the counterfactual-dependence vignette or to the no-counterfactual-dependence vignette. Of the 103 assigned to the dependence vignette, 71 said that Greg was a cause of Smith winning (68.9%); whereas, of the 93 assigned to the no-dependence vignette, 33 said that Greg was a cause of Smith winning (35.5%). A chi-square test of proportions shows that the proportion of "yes" answers was statistically significantly different in the two conditions: $\chi^2 = 20.63$, df = 1, p = 5.6e-6. According to Cohen's h, the effect size is given by h = arcsin(0.689) – arcsin(0.355) = 0.397, which would ordinarily be classified as somewhere between a small and a medium effect. Both proportions were also statistically different from chance, though in different directions.

[v] Purely structural accounts of actual causation must treat isomorphic causal structures the same way. Chockler and Halpern's account as well as Gerstenberg and Lagnado's development are purely structural, and the Overdetermination and Bogus Prevention cases appear to be isomorphic. And yet, the Overdetermination and Bogus Prevention cases appear to elicit different judgments. If the cases are really isomorphic and if ordinary people really say different things about the two, then the accounts are deficient. No data has been gathered on this question as far as we know. Moreover, Blanchard and Schaffer (forthcoming) argue that the simple model on which Overdetermination and Bogus Prevention appear to be isomorphic is not apt and that differences in the judgments elicited by Overdetermination and Bogus Prevention may be explained by differences in the structural models that are appropriate for the two cases.

[vi] What makes a choice of causal model or default values the *right* one depends on one's goals and on one's attitudes toward psychological models. For example, if the goal is to say how the cognitive mechanism works, the right causal model and default values will need to reflect representations that people actually have, but if the goal is more instrumental, the right causal model and default values might just be the ones that let a researcher reliably predict behavior.

[vii] An interesting alternative suggested by Knobe (personal communication) is that deviant status might be something that comes in degrees, and that people might regard an event as more causal to the extent that it is deviant. Halpern and Hitchcock (forthcoming) provide one framework for integrating graded causality, norms, and pivotality.

[viii] Hitchcock and Knobe were well-aware of previous work on the relation between abnormality and causation. They explicitly discuss Hart and Honoré, Kahneman and Miller, Hilton, and others in their paper.

[ix] Population-level statistical norms are statistical norms relative to a group of individuals. For instance, it's a population-level statistical norm that two year old children don't smoke: statistically speaking, two year old children, as a group, tend not to smoke cigarettes. In contrast, agent-level statistical norms are statistical norms relative to a particular agent's pattern of behavior. For instance, Aldi is a two year old who smokes every day (http://abcnews.go.com/Health/smoking-baby-today/story?id=14453373). While Aldi's smoking a cigarette is abnormal at the population level, it's a normal behavior *for him*.

[x] For ease, we are only discussing the role of blame in causal judgment. But, as Alicke, Rose, and Bloom (2011) argue, causal assessments can also be influenced by a desire to *praise*.