Following the FAD:
Folk Attributions and Theories of Actual Causation

Jonathan Livengood, Justin Sytsma, and David Rose

Abstract. In the last decade, several researchers have proposed theories of actual causation that make use of structural equations and directed graphs. Many of these researchers are committed to a widely-endorsed *folk attribution desideratum* (FAD), according to which an important constraint on the acceptability of a theory of actual causation is agreement between the deliverances of the theory with respect to specific cases and the reports of untutored individuals about those same cases. In the present article, we consider a small collection of related theories of actual causation: the purely structural theory developed in Halpern and Pearl (2005), and two theories that supplement the structural equations with considerations of defaults, typicality, and normality—Hitchcock (2007a) and Halpern and Hitchcock (2015). We argue that each of these three theories are meant to satisfy the FAD, then present empirical evidence that they fail to do so for several variations on a simple scenario from the literature. Drawing on the *responsibility view* of folk causal attributions suggested by Sytsma, Livengood, and Rose (2012), we conclude by offering a solution that allows the latter two theories to satisfy the FAD for these cases. The solution is to give up on concerns with typicality and focus on injunctive norms in supplementing the graphical modeling machinery.

Imagine a trolley beginning its descent down a steep hill on a rainy night. Before it begins to move, the brake operator says to the conductor of the trolley, "The cable has come loose, so if we need to slow down on the descent, we will have to rely exclusively on the handbrake." The conductor decides to proceed anyway. Halfway down the hill, the brake fails, and the out-of-control trolley crashes. Suppose that the trolley would not have crashed if it had not been raining, that it would not have crashed if the conductor had taken time to reattach the cable, and that it would not have crashed if the conductor had decided not to proceed at all. If those are the facts of counterfactual dependence in this case, what caused the trolley to crash? This is a question about *actual causation*, a question about which factor(s) from amongst the host of possible factors brought about the crash in the circumstances that actually obtained.

Many researchers working on actual causation have taken questions like "Did the rain cause the trolley to crash?" to be questions about the deliverances of common sense or ordinary intuition—what we'll refer to as *folk causal attributions*.[1] These researchers are committed to a *folk attribution desideratum* (FAD), according to which an important measure of the acceptability of a theory of actual causation is the agreement between its deliverances with regard to specific cases and folk causal attributions about those same cases.[2]

Over the past decade, many of these same researchers have employed the technical machinery of graphical causal modeling to provide an account of actual causation. Early accounts, like that of Halpern and Pearl (2005), were purely structural in character. Empirical evidence suggests, however, that normative considerations have a notable effect on folk causal attributions, rendering purely structural accounts a poor fit for the FAD.[3] Subsequent accounts, like those of Hitchcock (2007a) and Halpern and Hitchcock (2015), have fared better by supplementing the graphical modeling machinery with a distinction between default and deviant values of variables or with some consideration of normality.[4] Following Halpern and Hitchcock (2015, 3) we'll refer to such accounts collectively as DTN accounts (for defaults, typicality, and normality).

We hold that the shift toward DTN accounts moves theories of actual causation in the right direction for purposes of satisfying the FAD, and in this paper, we will provide new

[1] We will restrict talk of "folk causal attributions" to judgments about causation in concrete cases, such as the trolley case given above.

[2] Some commitment to common sense, intuitions, or the like has been very common in work on the metaphysics of causation. Paul and Hall (2013) offer a notable dissent in their "Rule five," which admonishes us not to enshrine intuitions. However, in fleshing out what they mean, they clearly state that they take intuitions to be valuable. "We think it is important to take intuitions very seriously, and we will do so throughout this book, paying special attention to places where our intuitions are in tension, since we take intuitions to be important guides to what we think we know about ontological structure, and the existence of said tensions indicate the need for further analysis. But intuitions must be used with care" (41).

[3] See Livengood and Sytsma (ms) for a recent line of evidence indicating that purely structural accounts fail to satisfy the FAD.

[4] See Danks (2016) and Livengood and Rose (2016) for overviews of graphical causal modeling and of experimental work on causal attribution, respectively. Sytsma and Livengood (2015) categorize work on causal attribution that is constrained by the FAD as part of a descriptive program in experimental philosophy. See Pearl (2000), Hitchcock (2001), Woodward (2003), Glymour and Wimberly (2007), and Halpern (2008) for applications of structural techniques to the problem of actual causation. See Hall (2007), Glymour et al. (2010), and Sytsma and Livengood (ms) for critical appraisals of these theories.

experimental evidence supporting that assessment. However, we will also present evidence that current DTN accounts fail to satisfy the FAD for some simple cases. We argue that the DTN accounts we consider fail to satisfy the FAD because they have not given sufficient priority to the role of injunctive norms in ordinary causal judgments.

Here is how we will proceed. In Section 1, we briefly describe three representative theories of actual causation. In Section 2, we note that the theories under consideration were meant to satisfy the FAD, apply the theories to Knobe's (2006) Lauren and Jane case, and provide empirical evidence that the deliverances of the theories do *not* match folk causal attributions for that case. In Sections 3 and 4, we consider possible objections and buttress our initial findings with further empirical results. Finally, in Section 5, we offer a suggestion for modifying the accounts under consideration to better satisfy the FAD.

1. Three Theories of Actual Causation

To test current theories of actual causation against the FAD we need to determine both what the theories and what the folk say about concrete cases. To determine the former, we need to know how the technical machinery works. In this section we very briefly discuss the technical machinery for these three theories.

1.1 Preliminaries

The three theories we consider all make use of graphical causal models. For present purposes, a *causal model* is an ordered pair consisting of an ordered set $\boldsymbol{V}$ of variables and an ordered set $\boldsymbol{E}$ of functions such that for each variable V in $\boldsymbol{V}$, there is a unique E_V in $\boldsymbol{E}$ that determines the value of V given the values of the other variables in $\boldsymbol{V}$. The functions in $\boldsymbol{E}$ are sometimes called *structural equations*. It will sometimes be valuable to partition the set $\boldsymbol{V}$ into an ordered set $\boldsymbol{U}$ of *exogenous* variables, whose values are directly assigned, and an ordered set $\boldsymbol{W}$ of *endogenous* variables, whose values are determined by the structural equations once the values of the variables in $\boldsymbol{U}$ are given. We will use the term "context" to denote the ordered set $\boldsymbol{u}$ of values assigned to the exogenous variables.

A *graphical* causal model is a causal model together with a directed graph—called a *causal graph*—that represents the functional dependencies in the causal model. Variables in a causal model become vertices of the corresponding causal graph. For each variable V in $\boldsymbol{V}$, there is a directed edge from X into V if and only if X appears as an independent variable (with non-zero coefficient) in the structural equation E_V. For example, the causal model $\mathscr{M}_1$ consisting of the set $<A, B, C, D>$ of variables and the set $<A = 1, B = (1 - A), C = B, D = B>$ of Boolean structural equations is represented by the causal graph in Figure 1. Let $X \rightarrow Y$ denote that there is a directed edge from X to Y. If $X \rightarrow Y$, then we say that X is a *parent* of Y and that Y is a *child* of X. A *path* of length $n \geq 0$ from V_i to V_j is a sequence $\boldsymbol{S} = <V_{(1)}, V_{(2)}, \ldots, V_{(n+1)}>$ of vertices such that $V_i = V_{(1)}$, $V_j = V_{(n+1)}$, and for every pair $<V_{(k)}, V_{(k+1)}>$ of vertices in $\boldsymbol{S}$, $V_{(k)} \rightarrow V_{(k+1)}$.

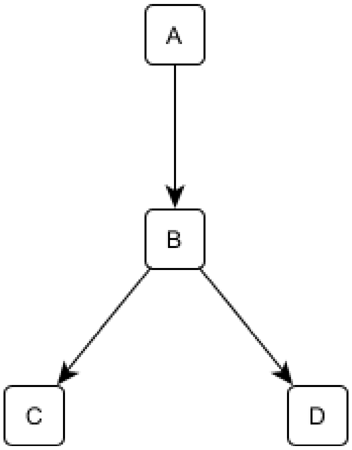

Figure 1: Causal graph of causal model $\mathscr{M}_1$.

1.2 Halpern and Pearl's Theory

In order to articulate Halpern and Pearl's theory, we need a small amount of additional notation. For any variable V in $\boldsymbol{V}$, if v is a possible value for V, then call $V = v$ a *primitive event*. Let φ denote a Boolean combination of primitive events, and let the expression $V \leftarrow v$ denote that the variable V has been *set* or *assigned* to the value v. For distinct variables $V_1, \ldots, V_n$, an expression of the form $[V_1 \leftarrow v_1, \ldots, V_n \leftarrow v_n]\ \varphi$ is called a *basic causal formula*.[5] A *causal formula* is a Boolean combination of basic causal formulas. If a causal formula ψ is true relative to a model $\mathcal{M}$ and a context $\boldsymbol{v}$, we write $(\mathcal{M}, \boldsymbol{v}) \vDash \psi$. Finally, for an ordered set $\boldsymbol{X}$ of variables $X_1, \ldots, X_n$ and an ordered set $\boldsymbol{x}$ of values $x_1, \ldots, x_n$, let $\boldsymbol{X} = \boldsymbol{x}$ denote the conjunction $X_1 = x_1 \wedge \ldots \wedge X_n = x_n$.

Halpern and Pearl offer the following account of actual causation (2005, 853).[6] Say that the event $\boldsymbol{X} = \boldsymbol{x}$ is an actual cause of the event ψ relative to $\mathcal{M}$ and $\boldsymbol{v}$ if and only if the following three conditions are all satisfied:

AC1. $(\mathcal{M}, \boldsymbol{v}) \vDash \boldsymbol{X} = \boldsymbol{x}$, and $(\mathcal{M}, \boldsymbol{v}) \vDash \psi$.

AC2. There is a partition of the endogenous variables $\boldsymbol{W}$ into sets $\boldsymbol{A}$ and $\boldsymbol{B}$ with $\boldsymbol{X} \subseteq \boldsymbol{A}$ and settings $\boldsymbol{x}' \neq \boldsymbol{x}$ and $\boldsymbol{b}'$ (which may or may not be equal to the actual values $\boldsymbol{b}$ of $\boldsymbol{B}$) such that if $(\mathcal{M}, \boldsymbol{v}) \vDash A = a$ for all A in $\boldsymbol{A}$, then both of the following are satisfied:

(a) $(\mathcal{M}, \boldsymbol{v}) \vDash [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{B} \leftarrow \boldsymbol{b}']\ \neg\psi$.
(b) $(\mathcal{M}, \boldsymbol{v}) \vDash [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{B}' \leftarrow \boldsymbol{b}', \boldsymbol{A}' \leftarrow \boldsymbol{a}]\ \psi$.[7]

AC3. $\boldsymbol{X}$ is minimal in the sense that no subset of $\boldsymbol{X}$ satisfies both AC1 and AC2.[8]

[5] Assuming that the v_i terms are possible values for the variables appearing in the basic causal formula.

[6] Halpern and Pearl talk about causal formulas as *events*, writing that "we are using the word 'event' here in the standard sense of 'set of possible worlds' (as opposed to 'transition between states of affairs'); essentially we are identifying events with propositions" (852, fn6). We will adopt their conventions here.

[7] The sets $\boldsymbol{A}'$ and $\boldsymbol{B}'$, which appear in condition **(b)**, are subsets of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, and condition **(b)** must hold for all such subsets.

[8] In Section 5 of their paper, Halpern and Pearl introduce a slight modification of their theory, but the added complication makes no difference with respect to the cases we consider. Hence, we will be concerned only with the original theory.

Condition (**AC1**) says that the cause and the effect both occur. Condition (**AC2**) says (in condition **a**) that a cause must be a difference-maker in some possibly counterfactual circumstance and (in condition **b**) that a cause has to actually do some difference-making work in the imagined circumstance. Condition (**AC3**) ensures that events are not counted as causes just because they have parts that are causes.

In order to make claims about actual causation, Halpern and Pearl's theory requires a causal model. In principle, the causal model can be discovered by appeal to experimental and statistical evidence. But in many cases like the ones we consider later on, the causal model is assumed to be obvious.[9]

1.3 Hitchcock's Theory

Hitchcock's theory consists of three parts: a causal model, a specification of the default values for the variables in the causal model, and a mathematical tool (**TC**).[10] To state **TC**, we need two new technical notions—a causal network and a self-contained causal network. A causal network is a subset of variables in a graphical causal model that satisfy a specific graphical condition:

> **CN.** Let <$\boldsymbol{V}$, $\boldsymbol{E}$> be a causal model, and let X, $Y \in \boldsymbol{V}$. The *causal network* connecting X to Y in <$\boldsymbol{V}$, $\boldsymbol{E}$> is 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 <$\boldsymbol{V}$, $\boldsymbol{E}$>. (509)

The notion of a self-contained causal network augments the graphical condition by appeal to the default values of the variables in the network:[11]

[9] However, see Halpern and Hitchcock (2010), Halpern (2015), and Blanchard and Schaffer (forthcoming) for discussion of why one ought to be careful in constructing a causal model.

[10] When the conditions specified by **TC** are not satisfied, the theory is silent about the actual causes. In what follows, we will restrict attention to cases that satisfy the conditions required for Hitchcock's theory to make actual causal attributions.

[11] Other authors, both within and without the graphical causal modeling tradition, have made use of a default/deviant distinction in order to try to answer questions about actual causation. For examples, see Hall (2007) and Halpern (2008). Blanchard and Schaffer (forthcoming) criticize the use of the default/deviant distinction in causal modeling.

> **SCN.** Let <***V***, ***E***> be a causal model, and let X, $Y \in \boldsymbol{V}$. Let $\boldsymbol{N} \subseteq \boldsymbol{V}$ be the causal network connecting X to Y in <***V***, ***E***>. Then the causal network ***N*** is *self-contained* if and only if for all $Z \in \boldsymbol{N}$, if Z has parents in ***N***, then Z takes a default value when all of its parents in ***N*** do (and its parents in $\boldsymbol{V} \setminus \boldsymbol{N}$ take their actual values). (510)

Suppose that the variables A, B, and C are all binary (0, 1), where zero is the default value. Consider the structural equation model $\mathcal{M}_2$ given by the Boolean equations $A = 1$, $B = 1$, and $C = A \wedge B$. In the model, both networks $\boldsymbol{N}_{AC}$ and $\boldsymbol{N}_{BC}$ are self-contained, since C would equal zero (its default) if either A or B were set equal to zero (their default values).

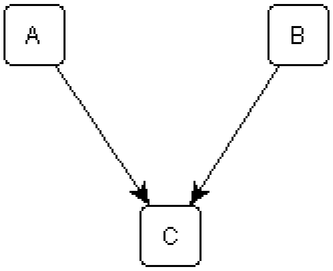

Figure 2: Causal graph for causal model $\mathcal{M}_2$.

With the notions of causal networks and default values of a variable in hand, Hitchcock produces **TC**, which says that in a self-contained network, counterfactual dependence of $Y = y$ on $X = x$ is both necessary and sufficient for $X = x$ to count as an actual cause of $Y = y$.

> **TC.** Let <***V***, ***E***> 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 <***V***, ***E***> is self-contained, then $X = x$ is a token cause of $Y = y$ in <***V***, ***E***> if and only if Y counterfactually depends upon X in <***V***, ***E***>. (511)

In other words, if the causal network connecting X to Y is self-contained, then $X = x$ is an actual cause of $Y = y$ if and only if for some non-actual value $x^* \neq x$ of variable X, variable Y takes on some non-actual value $y^* \neq y$.

Not surprisingly, Hitchcock holds that the difference between the default state of a system and deviations from that default is an important component of a correct theory of actual causation.[12] The difference, he claims, is fairly straightforward in most cases, but it is difficult to state precisely. Nonetheless, he offers a number of rules of thumb for determining the default and deviant values for a given case. Roughly, the default state of a system is its natural state—the state the system is (usually) in unless something has been done to it:

> As the name suggests, the default value of a variable is the one that we would expect in the absence of any information about intervening causes. More specifically, there are certain states of a system that are self-sustaining, that will persist in the absence of any causes other than the presence of the state itself: the default assumption is that a system, once it is in such a state, will persist in such a state. (2007a, 506)

In contrast, deviant states of a system are those that diverge from its natural state, and the system reaching that state typically involves an intervening cause:

> Temporary actions or events tend to be regarded as deviant outcomes. In the case of human actions, we tend to think of those states requiring voluntary bodily motion as deviants and those compatible with lack of motion as defaults. In addition, we typically feel that deviant outcomes are in need of explanation, whereas default outcomes are not necessarily in need of explanation. Frequently, but not always, my deviant values correspond to positive events, and defaults correspond to absences or omissions. (507)

It should be clear from these descriptions that Hitchcock's rules of thumbs do not offer a precise guide for determining default and deviant values for any given case. At the same time, the guidance he does offer needs to be taken seriously under threat of underdetermination. The danger is that by allowing the choice of default and deviant values to vary too widely, **TC** could be used to produce opposing causal attributions. That said, in what follows, we attempt to apply Hitchcock's rules of thumb charitably.

[12] He is not alone. See Hall (2007) for another take on the notion of defaults.

1.4 Halpern and Hitchcock's Theory

At the beginning of their essay on "Cause and Norm," Hitchcock and Knobe (2009) reflect on the problem of causal preemption and on attempts to handle the problem via broadly counterfactual approaches to actual causation:

> One promising line is to identify causation not with counterfactual dependence in the actual situation, but rather with counterfactual dependence in a certain kind of 'normalized' version of the actual situation. This normalized situation is reached by replacing abnormal features of the actual situation with more normal alternatives. (589)

Hitchcock and Knobe go on to argue that judgments of overall normality guide agents in choosing which counterfactuals to evaluate, and the choice of which counterfactuals to evaluate matters for causal attributions.

Halpern and Hitchcock (2015) make the role of normality in actual causation judgments precise by modifying condition **AC2a** in the Halpern and Pearl theory to require that the assignment of the values $\boldsymbol{x}'$ to $\boldsymbol{X}$ and $\boldsymbol{b}'$ to $\boldsymbol{B}$ in context $\boldsymbol{v}$ that makes ψ false yields a world that is at least as normal as the actual world. In order for an event to count as an actual cause of ψ, it cannot be strictly more normal than alternative events for which ψ is false.

The idea is best understood via an example. Consider the distinction between causes and background conditions.[13] Suppose an arsonist sets a fire in an abandoned warehouse, burning it to the ground. If the arsonist had not set a fire, the warehouse would not have burned down. But also, if there had been no oxygen in or around the warehouse at that time, the warehouse would not have burned down. In terms of counterfactual dependence, the arsonist and the oxygen are on par. However, while we are inclined to say that the arsonist caused the warehouse to burn down, we are not inclined to say that the oxygen caused the warehouse to burn down. What explains the difference? Halpern and Hitchcock suggest that while the world in which the arsonist *does not*

[13] This distinction is discussed in Section 7.3 of Halpern and Hitchcock (2015).

set a fire is more normal than the actual world (where she does), worlds in which there is no oxygen in the warehouse are clearly *less* normal than the actual world. In other circumstances, the presence of oxygen might count as a cause. If, for example, there is an oxygen leak in a vacuum chamber in a laboratory, the presence of oxygen might count as abnormal.

Following Hitchcock and Knobe (2009), Halpern and Hitchcock note two senses of "normal," writing:

> The word 'normal' is interestingly ambiguous. It seems to have both a descriptive and a prescriptive dimension. To say that something is normal in the descriptive sense is to say that it is the statistical mode or mean (or close to it). On the other hand, we often use the shorter form 'norm' in a more prescriptive sense. To conform with a norm is to follow a prescriptive rule. (429-430)

And while they hold that "further empirical research should reveal in greater detail just what kinds of factors can influence judgments of actual causation" (432), they also make it clear that they expect both types of norms—what they term *statistical norms* and *prescriptive norms*—to play a role in assessing normality.[14] In fact, they argue that "the different kinds of norm often serve as heuristic substitutes for one another" (430) and talk about "the extent to which we find it natural to glide between the different senses of 'norm'" (431). As with the rules of thumb in Hitchcock (2007a), it should be rather clear that Halpern and Hitchcock do not offer a precise guide for assessing normality. Nonetheless, their suggestions are meant to put some constraints on assignment of norms (2015, 433), as is needed to avoid the specter of underdetermination.

2. Lauren, Jane, the Theories, and the Folk

Many researchers interested in actual causation are committed to the FAD, according to which the deliverances of a theory of actual causation are correct insofar as they agree with folk causal

[14] Elsewhere we've referred to these under the labels of "descriptive norms" and "injunctive norms," taking the latter to include both prescriptions and proscriptions. For present purposes, however, we'll use Hitchcock and Knobe's terminology.

attributions. And this includes the three theories discussed in the previous section, as is made clear in the supplemental materials for this article. Accepting this, we want to know whether these theories satisfy the FAD. In this section, we test this for a test case due to Knobe (2006): Suppose that Lauren and Jane work at a company with a computer system that crashes if two or more people are logged in at the same time. The company knows how the system works and has instituted a policy governing how employees use the system. Lauren and Jane each log into the system and it crashes. In the circumstances, Lauren was permitted to log into the system, while Jane was not.

2.1 What the Theories Say

The simplest model that seems to capture the relevant details is one in which variables are specified for Lauren, Jane, and the state of the computer system (but see 3.1). Let *L*, *J*, and *C* be variables representing Lauren, Jane, and the computer system, respectively. The possible values of the variables *L* and *J* are "logs in" and "does not log in." The possible values for *C* are "crashed" and "not crashed." The equations assign values to both *L* and *J* directly. The variable *C* takes the value "crashed" if and only if both Lauren and Jane take the value "logs in." Since *L* and *J* are both assigned the value "logs in," *C* takes the value "crashed." Structurally, the model is identical to $\mathcal{M}_2$ (see 1.3).

Taking Halpern and Pearl's theory to make predictions about folk causal attributions, it predicts that when people are presented with the Lauren and Jane case they will say both that Lauren caused the system to crash and that Jane caused the system to crash. Hitchcock's theory makes the same prediction if we take the default values for Lauren and Jane to be "does not log in" and the default value for the computer system to be "not-crashed." We think these choices follow from charitable application of Hitchcock’s rules of thumb. First, not logging into the computer system is a *self-sustaining absence*, while logging in is a *temporary action*, a *positive event*, and one *requiring voluntary bodily motion*. Second, while the computer system continuing to run is not an absence, it is *self-sustaining* in the sense that if a computer is running, we generally expect it to continue running unless something disrupts it. Treating the computer in this way is similar to the way Hitchcock treats being alive for an ordinary human (2007a, 506).

Under some reasonable assumptions, Halpern and Hitchcock’s theory also predicts that people will say that both Lauren and Jane caused the crash. Starting with Jane, since she violates the company policy by logging in, it seems that a world in which she does not log in (with all else staying the same) is strictly more normal than the actual world. For Lauren, we think that it is most natural to say that her not logging in is neither more nor less normal than her logging in. Recall that in their discussion of normality, Halpern and Hitchcock distinguish between statistical norms and prescriptive norms. Let’s begin with prescriptive norms. Unlike Jane, Lauren is permitted to log in, so she does not violate a prescriptive norm by doing so. At the same time, being permitted to log in does not mean that she is required to do so and the story does not indicate that logging in is required for her usual work. As such, with regard to the prescriptive norm, it seems that Lauren not logging in is neither more nor less normal than Lauren logging in. Similarly for the statistical norm. The story does not specify that Lauren typically logs in, nor does it describe logging in as part of Lauren’s daily routine. If anything, we might expect that it is surely the case that logging in is more the exception than the rule for Lauren. Putting the norms together, it seems that Lauren not logging in is at least as normal as Lauren logging in.

Halpern and Hitchcock (2015, 436) consider the possibility of using normality orderings on worlds (called *witnesses*) that satisfy **AC2a** in order to grade or rank actual causes.[15] Suppose that there are several witnesses that $\boldsymbol{X} = \boldsymbol{x}$ is an actual cause of φ. Then we can consider the best witnesses, where a witness *s* is a *best* witness if there is no strictly more normal witness (i.e. there is no world that satisfies **AC2a** and is strictly more normal than *s*). Halpern and Hitchcock suggest ranking actual causes according to the relative normality of their best witnesses, and they go on to make the following empirical conjecture:

> We expect that someone's willingness to judge that $\boldsymbol{X} = \boldsymbol{x}$ is an actual cause of φ increases as a function of the normality of the best witness for $\boldsymbol{X} = \boldsymbol{x}$ in comparison with the best witness for other candidate causes. Thus, we are less inclined to judge that $\boldsymbol{X} = \boldsymbol{x}$ is an actual cause of φ when there are other candidate causes of equal or higher rank (436).

One might try to explain some of our empirical results by ranking actual causes in this way (e.g., the results of Studies 1 and 2). We are not in position to fully evaluate the proposal, since Halpern and Hitchcock do not make any guesses about the *degree* to which ordinary attributions of actual causation might be affected by the relative normality of the best witnesses for some collection of actual causes. However, we think that some of our empirical results do not sit comfortably with the proposal (e.g., the results of Studies 9 and 10).

2.2 What the Folk Say

What do ordinary people say about Knobe's Lauren and Jane case? To find out, we gave participants a vignette based on Knobe's thought experiment.[16] Participants were asked to indicate how strongly they agreed or disagreed with each of the two claims below on a 7-point

[15] We would like to thank an anonymous referee for reminding us about this part of Halpern and Hitchcock's paper.

[16] The vignettes for each of the studies in this article are provided in the supplemental materials.

Likert scale anchored at 1 with "strongly disagree," at 4 with "neutral," and at 7 with "strongly agree" (this scale was used in all of the studies reported in this article):

1. Lauren caused the system to crash.
2. Jane caused the system to crash.

As we expected, but in contrast to the causal attributions made by the theories, participants treated Lauren and Jane differently: they tended to say that Jane, but not Lauren, caused the system to crash.[17, 18] In a follow-up study, we replicated the finding of our first study using a between-participants design.[19] Each participant was given the same vignette as in our first study, but this time they were asked about just one of the two claims. Consistent with the results of our first study, participants tended to say that Jane, but not Lauren, caused the system to crash.[20] The results of are shown in Figure 3. Accepting these results, the three theories fail to satisfy the FAD for the Lauren and Jane case.[21]

[17] In all studies, responses were collected online through philosophicalpersonality.com; participants were native English speakers, 18 years of age or older, with at most minimal training in philosophy. Minimal training in philosophy was taken to exclude philosophy majors, those who have completed a degree with a major in philosophy, and those who have taken graduate-level courses in philosophy.

[18] N=72; 73.6% female, average age 34.6 years, ranging from 18-81 years old. In each study in this article we conducted one sample t-tests to compare the mean response for each claim to the neutral point of 4. Each test is one-tailed unless specified otherwise. For Study 1, the mean response for Lauren was significantly below the neutral point (mean=2.42, sd=2.04, t=-6.59, p=3.35e^{-9}) while the mean response for Jane was significantly above the neutral point (mean=5.21, sd=2.19, t=4.67, p=6.87e^{-6}).

[19] We want to thank an anonymous referee from *Review of Philosophy and Psychology* for suggesting this addition.

[20] N=34, 35; 65.2% female, average age 30.1, ranging from 18-67. Lauren: mean=2.41, sd=1.71, t=-6.59, p=3.35e^{-9}. Jane: mean=4.94, sd=2.29, t=4.67, p=6.87e^{-6}.

[21] An anonymous referee expressed a generic worry about our interpretation of our results. After remarking that our analysis assumes that responses lower than the midpoint express the judgment that the target factor is not a cause, the referee urged that people selecting 2 or 3 on our scale might be intending to say that the factor *is* a cause but not a very *strong* cause. But the referee's suggested interpretation of our data is implausible for two reasons. First, we explicitly asked participants to rate their level of agreement *or disagreement* with a statement, rather than asking them to rate the strength of a cause. Second, since we anchored 4 with the label "neutral" and 1 with "strongly disagree," the most natural reading of responses would understand responses of 2 or 3 to be disagreement—though admittedly less strong disagreement—with the target statement.

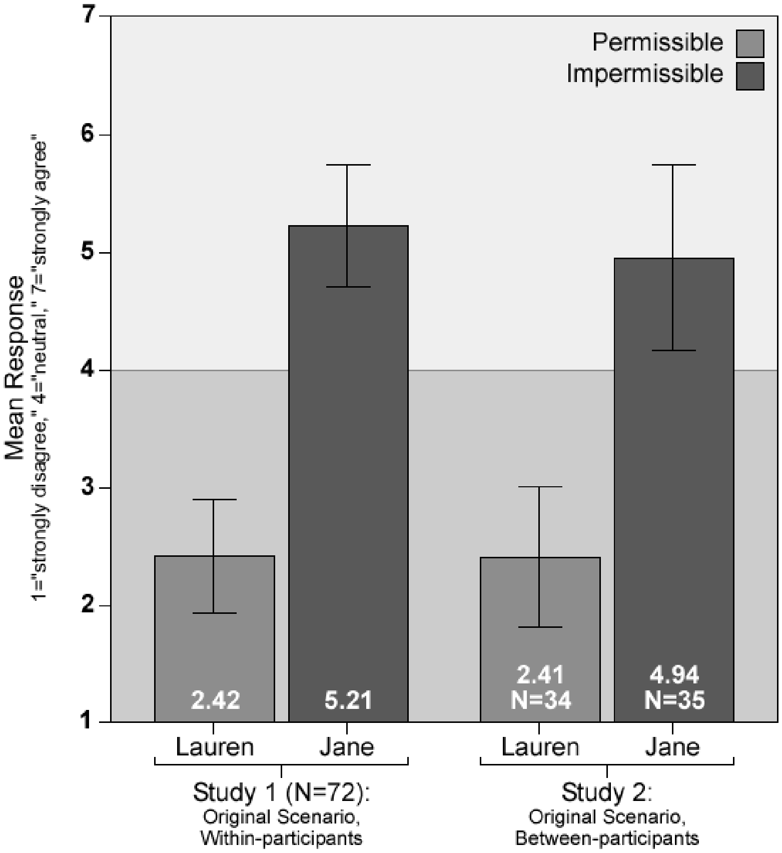

Figure 3: Results for Studies 1 and 2.

3. Objections and Replies

Our first two studies present prima facie reason to think that the theories of actual causation considered here do not satisfy the FAD. We expect two basic kinds of objection. The first kind aims to defend both purely structural theories and DTN theories, the second kind only DTN theories. We consider the first category in this section and the second in Section 4.

3.1 Objection 1: Wrong Causal Model

As noted in Section 1, none of the theories we are considering offers much guidance with regard to selecting a causal model. In testing the theories, we chose the simplest model that captured what we took to be the relevant details from the Lauren and Jane scenario. However, one might argue that the causal model we selected leaves out an important variable—the state of the computer system. Such a worry is especially pressing, since as pointed out by Halpern and Hitchcock (2010) and Halpern (2015), what a theory of actual causation says depends crucially on details about the causal model.

Adding a fourth variable to the causal model to capture the state of the mainframe, however, does not change what Halpern and Pearl's theory says *about Lauren and Jane*. Since the final state of the computer system still depends on what Lauren does and on what Jane does in the actual situation, Halpern and Pearl's theory counts both of them as actual causes of the crash. And the same holds for the other two theories. So this objection on its own does not enable the theories to satisfy the FAD for the Lauren and Jane case.

4.2 Objection 2: The Saliency of the Instability

Although simply changing the causal model does not save the theories we are considering, one might charge that our studies did not adequately test folk causal attributions with regard to the updated causal model. Specifically, one might argue that the instability of the mainframe is the most salient causal factor in the Lauren and Jane case but that we downplayed the instability by not asking a question about it, potentially skewing participants' responses for the statements about Lauren and Jane.

We tested the saliency objection in three ways. First, in Study 3, we gave participants the same probe used in our first study, but we also asked participants to rate how strongly they agreed or disagreed with a claim about the mainframe. In contrast to the causal attributions made by the theories, but in line with the results of our previous studies, participants tended to disagree with the claim that Lauren caused the system to crash but tended to agree with the claim that Jane caused the system to crash. In addition, they tended to agree with the claim that the mainframe caused the system to crash.[22] Second, in Study 4 we extended the between-participants design from our second study by asking participants a question about the mainframe. Interestingly, the mean response dropped significantly, with participants tending to deny that the mainframe caused the crash.[23] This suggests that the instability is not the most salient causal factor in the case, which undermines the saliency objection. Third, in Study 5 we rewrote the vignette to make the instability a feature rather than a bug. We specified that the company's mainframe is running an operating system that is *designed* to support only a single user at a time. Once again participants tended to deny that Lauren caused the crash and to affirm that Jane caused the crash.[24] We replicated Study 5 using a between-participants design in Study 6.[25] The results of these studies are shown in Figure 4. Taken together, they provide strong evidence that the instability of the computer system is not driving participants' responses with regard to Lauren and Jane.

[22] N=54; 76.9% female, average age 38.8, ranging from 18-75. Lauren: mean=1.54, sd=1.16, t=-15.59, $p<2.2e^{-16}$. Jane: mean=4.98, sd=2.29, t=3.14, p=0.00137. Mainframe: mean=5.24, sd=2.05, t=4.46, $p=2.18e^{-5}$.

[23] N=47; 66.0% female, average age 35.3, ranging from 18-61. Mainframe: mean=3.17, sd=2.01, t=-2.82, p=0.00699, two-tailed. Compared to corresponding question in Study 3: t(97.483)=-5.12, $p=1.57e^{-6}$.

[24] N=52; 75.5% female, average age 37.4, ranging from 18-74. Lauren: mean=2.08, sd=1.83, t=-7.56, $p=3.59e^{-10}$. Jane: mean=5.44, sd=2.11, t=4.93, $p=4.53e^{-6}$.

[25] N=36, 35; 70.4% female, average age 35.2, ranging from 18-74. Lauren: mean=1.72, sd=1.37, t=-10.01, $p=4.11e^{-12}$. Jane: mean=4.86, sd=2.38, t=2.1323, p=0.0201.

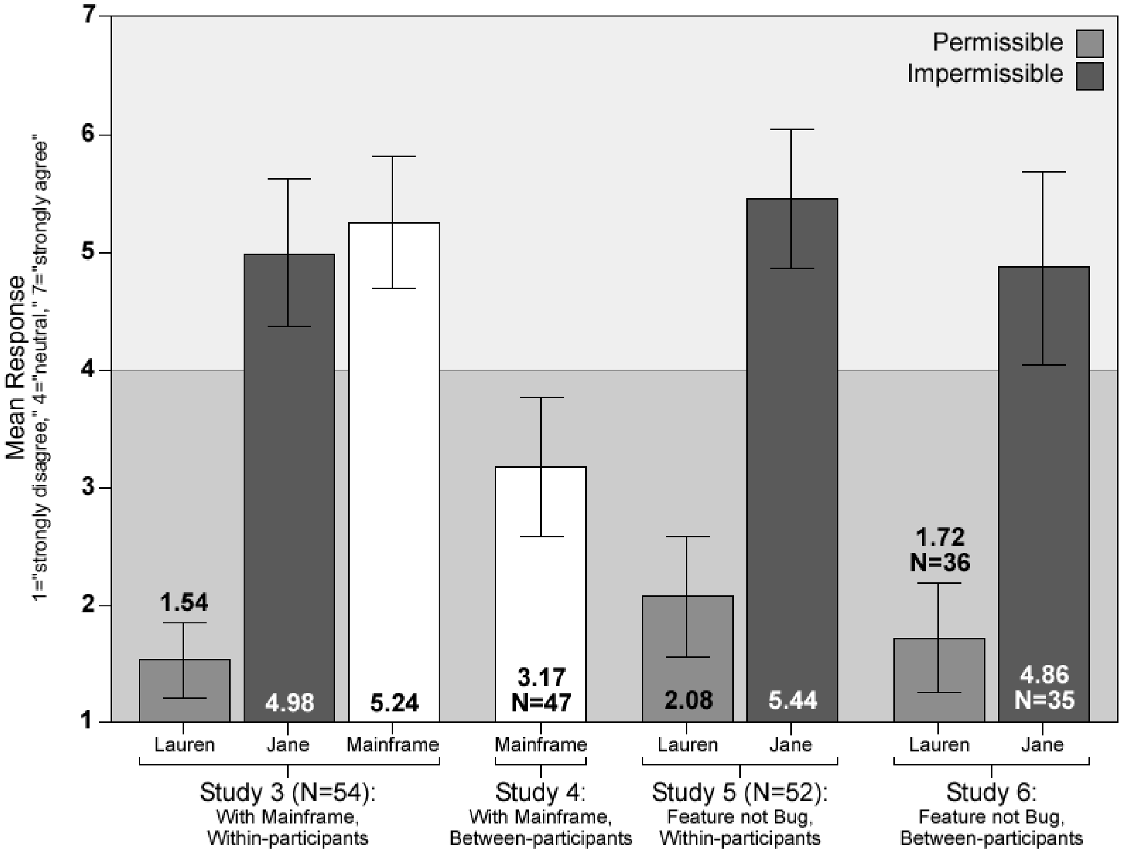

Figure 4: Results for Studies 3–6.

3.3 Objection 3: Blame Validation

One might argue that participants' reports do not reflect their (correct) causal *intuitions*. Perhaps people think that Jane deserves to be blamed (having violated company policy) but that Lauren should *not* be blamed (having followed the policy). In order to validate their blame judgments, people say that Jane but not Lauren caused the system to crash despite having the intuition that both Lauren and Jane caused the system to crash.[26] If so, it could be argued that their causal attributions are biased and should not constrain theorizing about actual causation.

[26] See Alicke (1992) and Alicke, Rose, and Bloom (2011) for a suggestion of this type; see Sytsma and Livengood (ms) for a response.

To test the blame validation objection, we carried out six further studies. In Study 7, we gave participants slightly different statements to evaluate after the original Lauren and Jane vignette. Instead of asking about the agents themselves, we asked participants to state their level of agreement with the following statements designed to emphasize the agents' actions:

1. Lauren's action of logging into the terminal caused the system to crash.
2. Jane's action of logging into the terminal caused the system to crash.

If participants' responses were being biased by blame judgments, we should be able to reduce or eliminate the effect by drawing attention away from the agents. However, participants denied that Lauren's action caused the crash and asserted that Jane's action caused the crash.[27] We replicated Study 7 using a between-participants design in Study 8.[28]

Perhaps focusing attention on the actions rather than the actors is inadequate. A more direct approach is to remove permissibility information from the Lauren and Jane scenario. Without permissibility information, the desire to blame should not bias ordinary causal attributions. What do the three theories we are considering predict for this case? Following the same logic articulated in Section 2, each of the theories would make the same predictions as they did for the original Lauren and Jane case.

In Study 9 we removed all information about the company's log-in policy from the Lauren and Jane scenario. Unlike in our previous studies, in this study participants treated Lauren and Jane equivalently; but contra the theories under consideration, our participants

[27] N=48; 66.7% female, average age 32.8, ranging from 18-59. Lauren: mean=2.52, sd=2.06, t=-4.97, p=4.68e^{-6}. Jane: mean=5.56, sd=2.08, t=5.20, p=2.14e^{-6}.

[28] N=43, 34; 68.8% female, average age 32.8, ranging from 18-64. Lauren: mean=2.51, sd=1.94, t=-5.02, p=4.99e^{-6}. Jane: mean=4.97, sd=2.18, t=2.59, p=0.0007.

tended to say that *neither* Lauren nor Jane caused the system to crash.[29] This study was replicated in Study 10 using a between-participants design.[30]

Finally, in Study 11 we combined the previous two approaches, giving participants a vignette with no permissibility information together with statements emphasizing actions as opposed to agents. Consistent with the previous results, participants denied both that Lauren's action caused the crash and that Jane's action caused the crash.[31] We replicated Study 11 using a between-participants design in Study 12. The results for Studies 7–12 are shown in Figure 5.

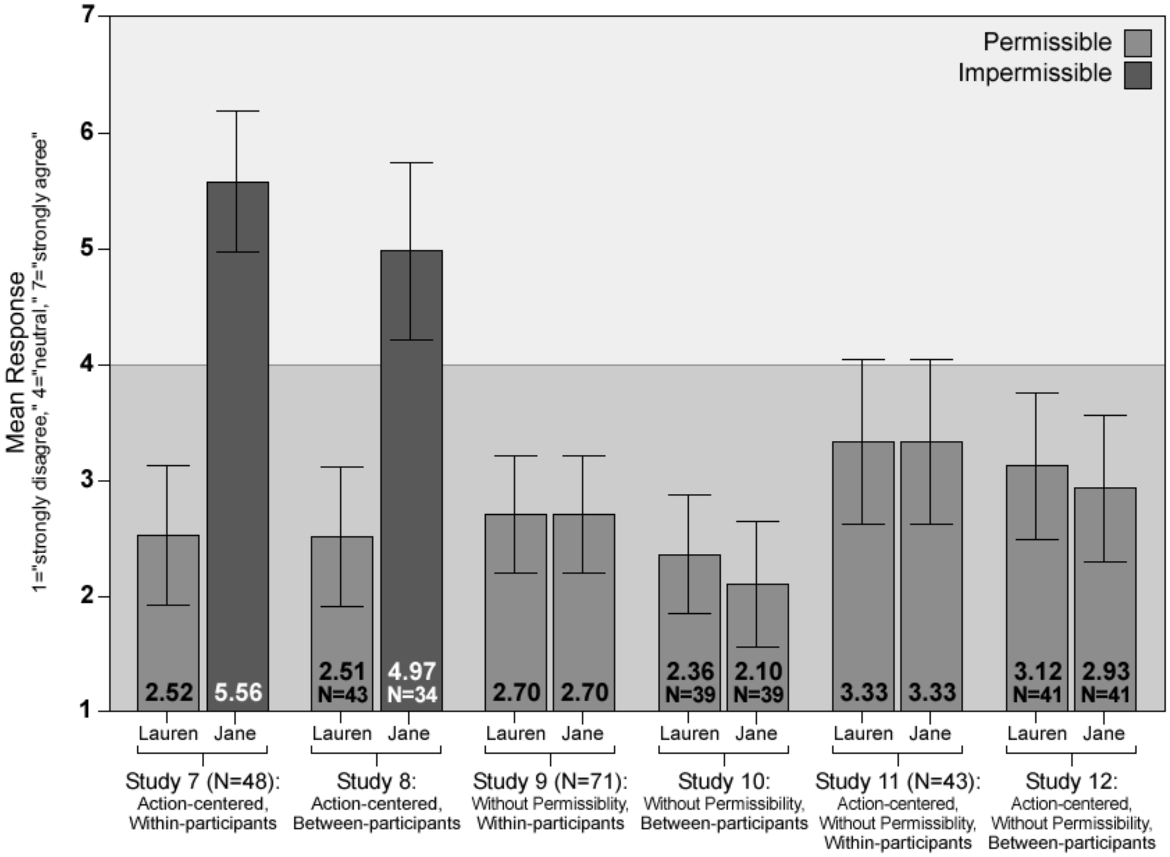

Figure 5: Results for Studies 7–12.

[29] N=71; 70.0% female, average age 35.1, ranging from 18-64. Lauren: mean=2.70, sd=2.14, t=-5.10, p=1.39e^{-6}. Responses for Jane were identical to those for Lauren.

[30] N=39, 39; 66.7% female, average age 39.3, ranging from 18-71. Lauren: mean=2.36, sd=1.56, t=-6.5518, p=5.00e^{-8}. Jane: mean=2.10, sd=1.67, t=-7.1085, p=8.77e^{-9}.

[31] N=43; 74.4% female, average age 30.0, ranging from 18-63. Lauren: mean=3.33, sd=2.31, t=-1.92, p=0.031. Responses for Jane were identical to those for Lauren.

4. Revisiting the Defaults and Norms

In Section 3, we responded to some attempts to defend both structural and DTN theories of actual causation from the criticism that they fail to satisfy the FAD in some simple cases. In this section we consider some defenses that are specific to DTN theories.

4.1 Solution 1: Focus on Permissibility

Let's focus on Hitchcock's theory for a moment. Hitchcock's theory fails to satisfy the FAD for the original Lauren and Jane case at least in part because the rules of thumb for assigning default and deviant values say nothing about prescriptive norms. Many studies have shown that ordinary causal attributions are sensitive to such norms (e.g., Alicke, 1992; Hitchcock and Knobe, 2009; Alicke et al., 2011; Sytsma, Livengood, and Rose, 2012). Since the Lauren and Jane case includes information about prescriptive norms, it is not completely surprising that Hitchcock's theory fails to satisfy the FAD with respect to it.

Is there a way to call on Hitchcock's rules of thumb to arrive at default and deviant values that take permissibility information into account? Allowing ourselves the benefit of hindsight, we could call on the thought that "we typically feel that deviant outcomes are in need of explanation" (2007a, 507), and argue that Jane's action—but not Lauren's—stands in need of explanation. Following this line of thought, we should treat the default value for Jane as being "does not log in" because Jane's logging in stands in need of explanation. Hence, Jane remains a cause. But what about Lauren? Having focused on the *impermissibility* of Jane's behavior, we might similarly focus on the *permissibility* of Lauren's behavior. We might assign Lauren the default value of "logs in" because her action does not call out for explanation. After all, she does not violate any company policy by logging in. In this way, Hitchcock's theory can be made to generate causal attributions that agree with the results from Study 1: Jane is said to be an actual cause of the system crashing, while Lauren is not. And a similar move can be made for Halpern and Hitchcock's theory.

By adjusting the defaults (or the normality rankings), the DTN theories can capture the asymmetry between the responses for Lauren and the responses for Jane. But, it should be noted that we could explain the asymmetry in two different ways: either the causal ratings for Jane are *elevated* relative to the causal ratings for Lauren, or the causal ratings for Lauren are *depressed* relative to the causal ratings for Jane. And these two explanations generate very different predictions for what we will find when we remove the permissibility information from the Lauren and Jane vignette. If the elevated explanation is correct, then people should tend to say that *neither* Lauren *nor* Jane caused the system to crash. And if the depressed explanation is correct, then people should tend to say that *both* Lauren *and* Jane caused the system to crash.

Using permissibility as a way to bring the predictions for the DTN theories in line with folk causal attributions for the original Lauren and Jane case aligns the theories with the *Lauren depressed* explanation. As such, the theories predict that if we remove permissibility information from the Lauren and Jane scenario, people will say both that Lauren caused the system to crash and that Jane caused the system to crash.[32] But we saw in Study 9 that when permissibility information is removed, participants tend to say that *neither* Lauren *nor* Jane caused the system to crash—a finding that was further supported by the remaining studies in Section 4. It appears,

[32] Hitchcock's theory was able to satisfy the FAD for the original case by treating the impermissibility of *Jane's* action as shifting the default value for *Lauren* from "does not log in" to "logs in." In the absence of permissibility information, however, no shift would occur. Hence, without permissibility information, our initial verdict follows from Hitchcock's rules of thumb. Not logging into the computer system is a self-sustaining absence. Logging into the system is a temporary action that requires voluntary bodily motion. Using the default value of "does not log in" for both Lauren and Jane, Hitchcock's theory again asserts that both are actual causes of the system crashing. Similarly, Halpern and Hitchcock's theory predicts that people will tend to say that both Lauren and Jane caused the system to crash when no permissibility information is given, provided they judge worlds in which Lauren (or Jane) logs in to be no more normal than worlds in which she does not log in.

then, that the *Jane elevated* explanation is the correct one. Thus, not only do the theories fail to satisfy the FAD for the Lauren and Jane case when permissibility information is removed, but insofar as they are able to satisfy the FAD for the original Lauren and Jane case, they get the causal attributions right for the wrong reason: they depress Lauren rather than elevating Jane.

4.2 Solution 2: Revisiting the Saliency of the Instability

Perhaps a variation on the objection from 3.2 would allow the DTN theories to handle cases where permissibility information is removed. Specifically, one might argue that ordinary people take the instability of the system to be *in need of explanation*, while the actions of Lauren and Jane are not. If so, then further modification of the default values or the normality rankings might enable the DTN theories to also satisfy the FAD for cases without permissibility information.

For Hitchcock's theory, the idea would be to focus on the rule that "we typically feel that deviant outcomes are in need of explanation." If only the instability of the computer system calls for explanation, we can set the default values for Lauren and Jane to be "logs in," and the default value for the computer system to be "stable." With those defaults, Hitchcock's theory correctly predicts folk responses to the Lauren and Jane case without permissibility information. A similar story will patch Halpern and Hitchcock's theory.

Hitchcock and Knobe (2009) distinguish between three types of norms. So far we have focused on statistical norms and prescriptive norms, but there are also norms of proper functioning. Halpern and Hitchcock also discuss norms of proper functioning, writing:

> There are specific ways that human hearts and car engines are 'supposed' to work, where 'supposed' here has not merely an epistemic force, but a kind of normative force. Of course, a car engine that does not work properly is not guilty of a moral wrong, but there is nonetheless a sense in which it fails to live up to a certain kind of standard. (430)

With regard to the Lauren and Jane without permissibility information scenario, it might then be urged that the most salient norm violation is that the instability of the mainframe violates a norm of proper functioning: mainframes are not *supposed* to crash when more than one person logs in. Relative to this abnormality, it might then be thought that Lauren and Jane logging in is comparably normal—perhaps to the point of overshadowing that the relevant comparison is between their logging in and their not logging in, leading people to judge that worlds in which Lauren/Jane log in are more normal than worlds in which they don't. Using these normality judgments, Halpern and Hitchcock's theory correctly predicts folk responses in the Lauren and Jane case when no permissibility information is provided.

The instability explanation solution fits well with what we find when we add a question about the mainframe to the probe used in Study 9. As predicted by the revised theories, in Study 13 we found that participants disagreed with both the statement that Lauren caused the crash and the statement that Jane caused the crash, but that they agreed with the statement that that the mainframe caused the crash.[33] We replicated this study using a between-participants design in Study 14.[34]

Recall that in 3.2, we tested a variation on the Lauren and Jane scenario in which the operating system was described as being *designed* to support only a single user at a time. We found that this revision did not have a notable effect on participants' responses. But what happens if we remove the permissibility information? Since in the new story, the mainframe is not behaving abnormally, people would have no reason to adopt the default values or normality rankings suggested for the version in which the system is behaving abnormally. Hence, both

[33] N=62; 67.7% female, average age 35.1, ranging from 18-80. Lauren: mean=2.74, sd=1.87, t=-5.29, p=8.74e^{-7}. Jane: mean=2.56, sd=1.76, t=-6.42, p=1.15e^{-8}. Mainframe: mean=6.44, sd=1.03, t=18.54, p<2.2e^{-16}.

[34] N=40, 47, 40; 70.9% female, average age 35.7, ranging from 18-80. Lauren: mean=1.95, sd=1.48, t=-8.74, p=5.08e^{-11}. Jane: mean=2.38, sd=1.69, t=-6.57, p=2.04e^{-8}. Mainframe: mean=4.85, sd=1.93, t=2.79, p=0.0041.

theories should again maintain that Lauren and Jane both caused the crash. We tested this in Study 15. Contrary to the predictions, participants tended to deny both that Lauren caused the system to crash and that Jane caused the system to crash.[35] We replicated this study using a between-participants design in Study 16.[36] The results are shown in Figure 6.

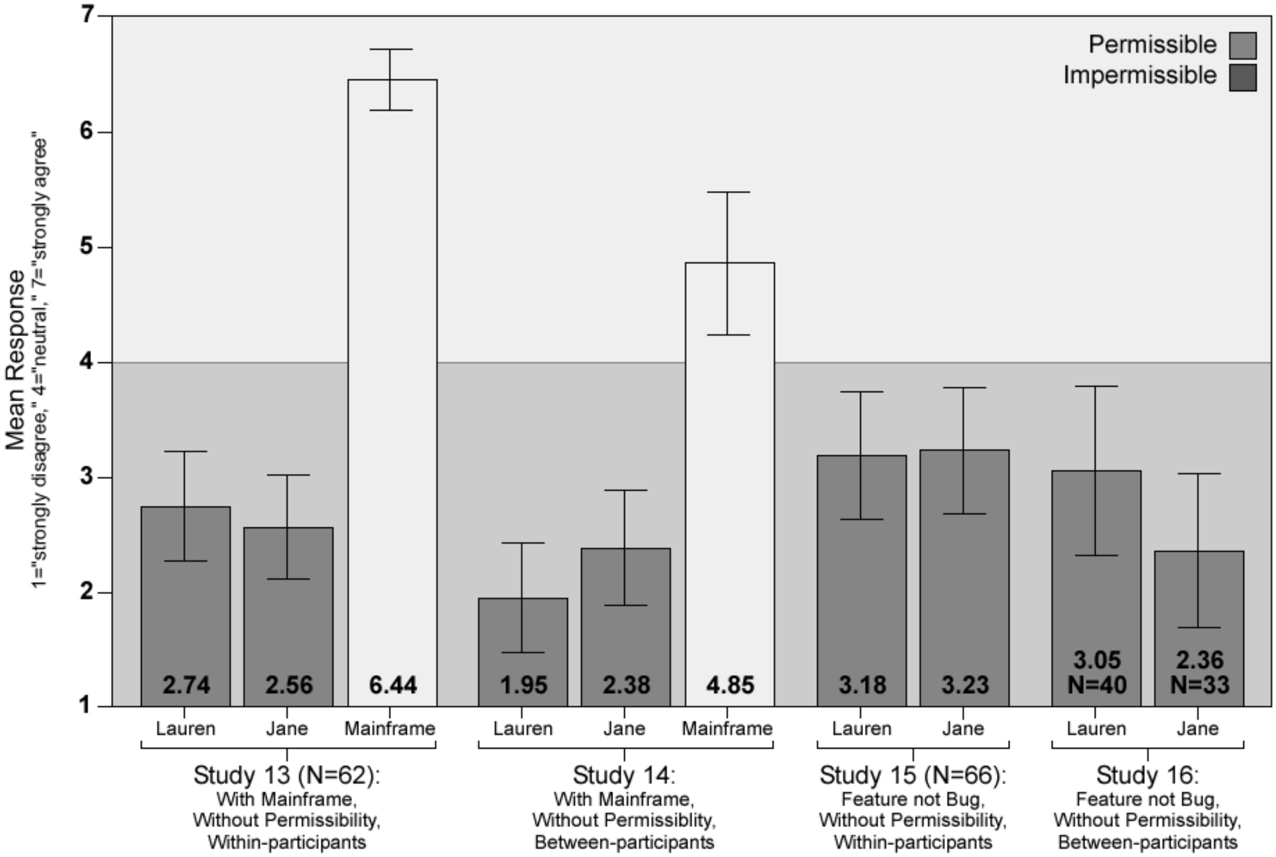

Figure 6: Results for Studies 13–16.

[35] N=66; 66.7% female, average age 37.1, ranging from 18-65. Lauren: mean=3.18, sd=2.24, t=-2.97, p=0.00209. Jane: mean=3.23, sd=2.21, t=-2.84, p=0.00301.

[36] N=40, 33; 65.8% female, average age 31.8, ranging from 18-58. Lauren: mean=3.05, sd=2.29, t=-2.63, p=0.00611. Jane: mean=2.36, sd=1.87, t=-5.03, p=9.03e^{-6}.

4.3 Solution 3: Focus on Typicality

In the previous two sub-sections we attempted to patch up the DTN accounts by focusing on two different types of norms—prescriptive norms in 4.1 and norms of proper functioning in 4.2—and while we made some progress in each case, the accounts still fail to satisfy the FAD for one of the variations on the Lauren and Jane case. Perhaps this can be rectified by calling on the third type of norm described by Hitchcock and Knobe (2009)—statistical norms.

Recall that in Section 2 we argued that because the Lauren and Jane story does not specify that Lauren typically logs in, and since we might expect that logging in is more the exception than the rule for her, her logging in is no more normal than her not logging in. However, one might argue that in the absence of other information about Lauren's job, people will assume that logging into the mainframe is a typical part of her work day. And the same could be argued for Jane in the cases where no permissibility information is given. If this is accurate, then Halpern and Hitchcock's theory would correctly predict that people will tend to deny that either Lauren or Jane caused the system to crash when they do not violate a company policy by logging in. Now, it might be responded that this is a stretch, especially when it comes to our last two studies: in Studies 15 and 16, the mainframe is described as being designed to support only a single user, such that it would seem highly doubtful that it would be both typical for Lauren to log in and typical for Jane to log in. That would, obviously, produce an unworkable situation.

However, we can test the objection more directly. In Sytsma, Livengood, and Rose (2012) we argued that there are two types of typicality that need to be considered when testing the effects of statistical norms on folk causal attributions—what is typical for an agent (agent-level typicality) and what is typical for members of the relevant population to which that agent belongs (population-level typicality).[37] In Study 17, we tested the role of both types of typicality on a further variation on the Lauren and Jane without permissibility information scenario. To simplify matters, we removed Jane from the story and specified that the system would crash if *anyone* logged in. Lauren was then said to log in, and that the system crashed. On a second page, participants were then given a follow-up vignette in which permissibility information was added.[38]

We ran five variations on this case. In the first variation, no typicality information was given. In the remaining four variations, Lauren's action was described as either agent-level typical, agent-level atypical, population-level typical, or population-level atypical respectively. Across these probes, we found that providing typicality information had no relevant effect: in each case participants tended to deny that Lauren caused the system to crash on the first page (no permissibility information) and to affirm that she caused the system to crash on the second page (permissibility information added).[39] The results are shown in Figure 7.

[37] Halpern and Hitchcock acknowledge this distinction in laying out their account (2015, 432).

[38] Participants were not able to change their response to the question on the first page.

[39] N=61, 50, 58, 45, 55; 78.8% female, average age 37.6, ranging from 18-77. Without typicality information: first page (mean=2.70, sd=2.02, t=5.01, p=$2.56e^{-6}$); second page (mean=5.56, sd=1.85, t(60)=6.58, p=$6.49e^{-9}$). Agent-level typical: mean=2.36, sd=1.80, t=-6.43, p=$2.56e^{-8}$; mean=5.92, sd=1.85, t=7.34, p=$9.91e^{-10}$. Agent-level atypical: mean=2.93, sd=2.07, t=-3.94, p=0.000114; mean=5.72, sd=1.90, t=6.91, p=$2.21e^{-9}$. Population-level typical: mean=2.89, sd=6.02, t=-3.32, p=0.000921; mean=6.02, sd=1.89, t=7.18, p=$3.1e^{-9}$. Population-level atypical: mean=2.96, sd=2.20, t=-3.49, p=0.000485; mean=5.82, sd=1.85, t=7.30, p=$6.68e^{-10}$.

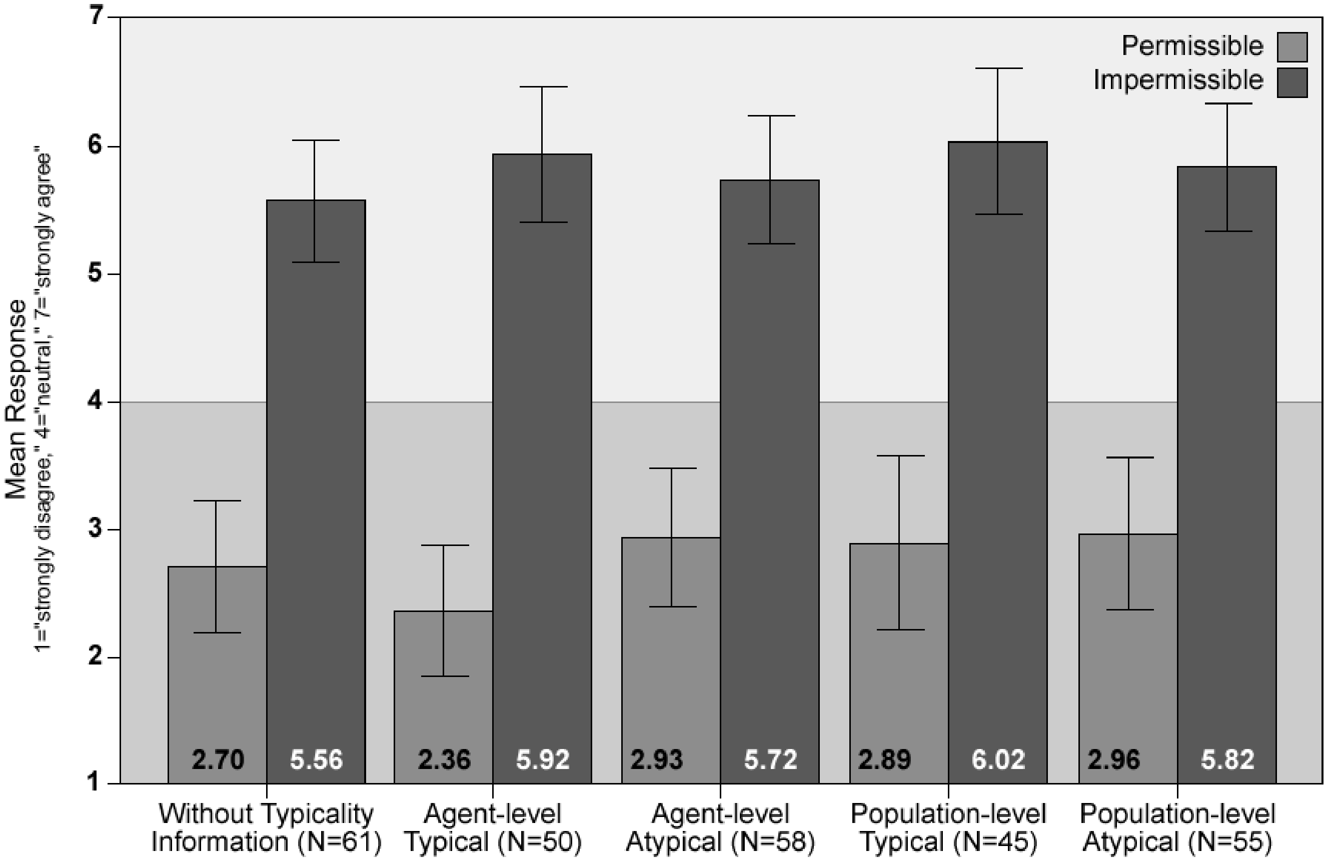

Figure 7: Results for Study 17.

5. Revising the DTN Accounts

DTN accounts have generally adopted a broad view of the normative factors involved in folk causal attributions. For instance, we've seen that Halpern and Hitchcock (2015) note three types of norms that they expect to be involved in normality judgments, including both statistical norms and prescriptive norms. In the previous section, however, we saw that statistical norms did not have a notable impact on folk causal attributions for the Lauren alone case. And this finding is in keeping with the results reported in Sytsma, Livengood, and Rose (2012).

Noting this, one might try slightly revising Halpern and Hitchcock's account to satisfy the FAD for the variations on the Lauren and Jane case we have considered by letting statistical norms do significantly less work in the account, either by removing talk of statistical norms entirely (since in the cases we've considered, statistical norms do not seem to matter for ordinary causal attributions) or by significantly altering the role of statistical norms in the account.[40] (A comparable move could be made with regard to determining default values for Hitchcock's theory.) Generalizing across the studies we've looked at in this article, Halpern and Hitchcock's theory runs into trouble with regard to satisfying the FAD in those situations where a prescriptive norm doesn't apply. And the reason is that when there isn't a prescriptive norm to call on, we've defaulted to calling on statistical norms that do not seem to matter to ordinary people.

[40] This does not mean that statistical norms should be ignored completely. Rather, we hold that what effects statistical norms have on folk causal attributions about agents occur by affecting judgments about prescriptive norms, which are already captured in the account. See Sytsma, Livengood, and Rose (2012) for discussion. That said, one might think that there is good reason to believe that statistical norms play an independent role in some folk causal attributions. An anonymous referee suggested two such reasons. First, it might be urged that *obviously* people do not hold non-agents responsible, such that statistical norms must play a role in folk causal attributions concerning non-agents that is independent of statistical norms. Here it should be noted that our focus in this paper is on understanding folk causal attributions with regard to agents, and the factors involved in attributions concerning non-agents might differ from those involved in attributions concerning agents. That said, we do not believe that it is obvious that people do not hold non-agents responsible. In our opinion this is an open empirical question that stands in need of testing. We tentatively hold, however, that our account will also cover folk causal attributions concerning non-agents. This expectation is based in part on preliminary testing indicating that responsibility judgments play a role in folk causal attributions. Further, background empirical work suggests that people tend to take an agentive perspective on nature as a whole (see e.g., Bloom, 2007; Kelemen, 2012; Rose, 2015; Rose and Schaffer, 2015), providing reason to expect that people will hold non-agents responsible. Second, it might be argued that the fourth experiment in Kominsky et al. (2015) provides empirical evidence suggesting that statistical norms play an independent role in folk causal attributions. In this study participants were presented with a scenario where Alex plays a game that involves both flipping a coin and rolling a pair of dice. In the relevant cases, for Alex to win he both needs to get a heads on the coin flip and get at least a certain number on the sum of the dice. In one case the sum needed is likely to occur (higher than 2) and in the other it is unlikely to occur (higher than 11). In each case Alex wins and participants are asked how strongly they agree or disagree with the statement, "Alex won because of the coin flip." Kominsky et al. found that participants gave significantly higher answers in the case where Alex only needed to roll higher than 2 than in the case where he needed to roll higher than 11. This is a fascinating result. It is unclear that it provides evidence for the role of statistical norms in folk causal attributions, however, because it is unclear that we should take agreement/disagreement with the statement "Alex won because of the coin flip" as indicating agreement/disagreement with a corresponding causal statement such as "the coin flip caused Alex to win." See Livengood and Machery (2007) for evidence that "X caused Y" and "Y because X" statements sometimes come apart. Further, we have tested this contrast for the first variation of the Lauren scenario used in Study 17. We asked participants whether "the system crashed because Lauren logged in" (as opposed to "Lauren caused the system to crash"). We found that judgments were significantly higher for the "because" statement than they were for the "caused" statement in Study 17: N=60, 70.0% female, average age 37.9, ranging from 18-71; mean=4.03, sd=2.48 (contrasted to 2.70, 2.02); t=-3.23, p=0.001645, two-tailed. It may also be the case that causal attributions are sensitive to certain varieties of statistical norm but not others. If so, then perhaps there is a salient difference between the statistical norms at play in the cases we've considered and the statistical norms at play in Kominsky et al.'s case. Further research on the role of statistical norms in causal attributions about cases like the one given by Kominsky et al. is called for and is currently underway.

This raises an important question: What considerations should guide normality judgments when the prescriptive norms allow an agent to either act or to refrain? Our suggestion follows from the responsibility view—i.e., the view that folk causal attributions are inherently normative and are closely related to responsibility judgments. With regard to assessing responsibility for a bad outcome (like the system crashing in the Lauren and Jane case), we expect that people tend to treat it as normal for the agent to do what they are allowed to do. The idea, here, is that it is acceptable for the agent to do what the prescriptive norms allow them to do, and because of this we shouldn't hold it against them if they thereby unwittingly bring about a bad outcome. If this is correct, then on the hypothesis that folk causal attributions correspond with responsibility judgments, we expect the same rule of thumb to work for arriving at normality judgments for purposes of predicting folk causal attributions.

Focusing on prescriptive norms and utilizing the rule of thumb given above, the DTN accounts we've been considering produce the correct predictions for each of the variations on the Lauren and Jane case we have looked at. Focusing on Halpern and Hitchcock's theory as applied to the original Lauren and Jane scenario, we see that Jane acts abnormally in logging in because she is prohibited from doing so, while Lauren acts normally in logging in because she is allowed to do so. And the result is that Halpern and Hitchcock's theory then correctly outputs the prediction that people will tend to say that Jane caused the system to crash and that Lauren did not cause the system to crash. Similarly for the cases where permissibility information is removed. In these cases our rule of thumb leads us to judge that both Lauren and Jane act normally in logging in because they are allowed to do so, which produces the correct prediction that people will tend to say that neither of them caused the system to crash.

6. Conclusion

Knobe’s (2006) Lauren and Jane case raises a serious problem for three prominent theories of actual causation. In opposition to the causal attributions delivered by these theories, participants tend to disagree with the statement that Lauren caused the system to crash. A range of follow-up studies testing various objections confirms the basic finding. And while the accounts calling on defaults, typicality, or normality can be patched-up to handle the original scenario, the resulting theories have problems with variations in which permissibility information is removed. The result is that the theories fail to satisfy the FAD and are in need of revision. We hold that the problem is that these theories place weight on statistical norms in addition to prescriptive norms, and that the most straightforward revision is to remove statistical norms from the account and replace them with additional guidance for navigating prescriptive norms. With such a revision, both Hitchcock’s (2007a) theory and Halpern and Hitchcock’s (2015) theory are able to satisfy the FAD for the cases we have considered.

References

Alicke, M. (1992). “Culpable Causation.” *Journal of Personality and Social Psychology*, 36: 368–378.

Alicke, M., D. Rose, and D. Bloom (2011). “Causation, Norm Violation and Culpable Control.” *Journal of Philosophy*, 108(12): 670-696.

Beebee, H. (2004). “Causing and Nothingness,” in *Causation and Counterfactuals*, edited by Collins et al., 291-308.

Bloom, P. (2007). “Religion is natural.” *Developmental Science*, 10, 147–151.

Bollen, K. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Blanchard, T. and J. Schaffer (forthcoming). “Cause without Default.” In *Making a Difference*, edited by H. Beebee, C. Hitchcock, and H. Price, Oxford University Press. Pre-print at: http://www.jonathanschaffer.org/cnod.pdf

Collins, J. et al., eds. (2004). *Causation and Counterfactuals*. Cambridge: MIT Press.

Danks, D. (2016). “Causal Search, Causal Modeling, and the Folk,” in *A Companion to Experimental Philosophy*, edited by J. Sytsma and W. Buckwalter, Blackwell.

Danks, D., D. Rose, and E. Machery (2014). “Demoralizing Causation.” *Philosophical Studies*, 171: 251-277.

Glymour, C. et al. (2010). “Actual Causation: A Stone Soup Essay,” *Synthese* 175, 169-192.

Glymour, C. and F. Wimberly (2007). “Actual Causes and Thought Experiments,” in *Causation and Explanation*, 43-67. Edited by Campbell, O’Rourke, and Silverstein. Cambridge: MIT Press.

Hacker, P. (2009). “Critical Studies: A Philosopher of Philosophy,” *Philosophical Quarterly* 59(235), 337-348.

Hall, N. (2007). “Structural Equations and Causation,” *Philosophical Studies* 132, 109-136.

Hall, N. (2004). “Rescued from the Rubbish Bin,” *Philosophy of Science* 71, 1107-1114.

Halpern, J. (2008). “Defaults and Normality in Causal Structures,” in *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference (KR ’08)*, 198-208.

Halpern, J. (2015). “Appropriate Causal Models and the Stability of Causation.” Unpublished manuscript. Arxiv preprint at: http://arxiv.org/pdf/1412.3518.pdf

Halpern, J. and C. Hitchcock (2010). “Actual causation and the art of modeling.” In Dechter, Geffner, and Halpern (Eds.) *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*, 383-406. London: College Publications.

Halpern, J. and C. Hitchcock (2015). “Graded Causation and Defaults,” British Journal for the Philosophy of Science, 66(2): 413-457.

Halpern, J. and J. Pearl (2005). “Causes and Explanations: A Structural-Model Approach. Part I: Causes,” *British Journal for the Philosophy of Science* 56, 843-887.

Hitchcock, C. (2009). “Structural equations and causation: six counterexamples,” *Philosophical Studies* 144, 391-401.

Hitchcock, C. (2007a). “Prevention, Preemptions, and the Principle of Sufficient Reason,” *Philosophical Review* 116(4), 495-531.

Hitchcock, C. (2007b). “Three Concepts of Causation,” *Philosophy Compass* 2/3, 508-516.

Hitchcock, C. (2001). “The Intransitivity of Causation Revealed in Equations and Graphs,” *The Journal of Philosophy* 98(6), 273-299.

Hitchcock, C. and J. Knobe (2009). “Cause and Norm.” *Journal of Philosophy* 106(11), 587-612.

Holland, P. (1986). “Statistics and Causal Inference,” *Journal of the American Statistical Association* 81, 945-960.

Ichikawa, J. (2009). “Explaining Away Intuitions,” *Studia Philosophica Estonica* 2(2), 94-116.

Kelemen,D. (2012). “Teleological minds: How natural intuitions about agency and purpose influence learning about evolution.” In K. S. Rosengren, S. K. Brem, E. M. Evans, & G. M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution*. Oxford: Oxford University Press.

Kline, R. (1998). *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.

Kominsky, J. et al. (2015). “Causal superseding.” *Cognition* 137, 196-209.

Knobe, J. (2006). *Folk Psychology, Folk Morality*. Dissertation.

Knobe, J. and B. Fraser (2008). “Causal judgments and moral judgment: Two experiments.” In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447. Cambridge: MIT Press.

Lewis, D. (2004). “Causation as Influence,” in *Causation and Counterfactuals*, edited by Collins et al., 75-106.

Lewis, D. (1986). *Philosophical Papers*, Volume II. Oxford: Oxford University Press.

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*(1), 107-127.

Livengood, J. and D. Rose (2016). “Experimental Philosophy and Causal Attribution,” in *A Companion to Experimental Philosophy*, edited by J. Sytsma and W. Buckwalter. Blackwell.

Livengood, J. and J. Sytsma (ms). “Actual Causation and Compositionality.”

Menzies, P. (1996). “Probabilistic Causation and the Pre-emption Problem,” *Mind* New Series 105(417), 85-117.

Paul, L. and N. Hall (2013). *Causation: A User's Guide*. Oxford: Oxford University Press.

Pearl, J. (2000). *Causality*: *Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Rose, D. (2015). “Persistence Though Function Preservation.” *Synthese*, 192, 97-146.

Rose, D. and J. Schaffer (forthcoming). “Folk Mereology is Teleological.” *Nous*.

Roxborough, C. and J. Cumby (2009). “Folk psychological concepts: Causation.” *Philosophical Psychology*, 22(2): 205–213.

Sytsma, J. and J. Livengood (2015). *The Theory and Practice of Experimental Philosophy*. Broadview.

Sytsma, J. and J. Livengood (ms). “Intervention, Bias, Responsibility... and the Trolley Problem.”

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: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4), 814-820.

Weatherson, B. (2003). “What Good are Counterexamples?” *Philosophical Studies* 115, 1-31.

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