Forthcoming in *Oxford Studies in Experimental Philosophy*

Causal Reasoning: Philosophy and Experiment[*]

James Woodward
History and Philosophy of Science
University of Pittsburgh

1. Introduction

The past few decades have seen much interesting work, both within and outside philosophy, on causation and allied issues. Some of the philosophy is understood by its practitioners as metaphysics—it asks such questions as what causation "is", whether there can be causation by absences, and so on. Other work reflects the interests of philosophers in the role that causation and causal reasoning play in science. Still other work, more computational in nature, focuses on problems of causal inference and learning, typically with substantial use of such formal apparatuses as Bayes nets (e.g. , Pearl, 2000, Spirtes et al, 2001) or hierarchical Bayesian models (e.g. Griffiths and Tenenbaum, 2009). All this work – the philosophical and the computational -- might be described as "theoretical" and it often has a strong "normative" component in the sense that it attempts to describe how we *ought* to learn, reason, and judge about causal relationships. Alongside this normative/theoretical work, there has also been a great deal of empirical work (hereafter Empirical Research on Causation or ERC) which focuses on how various subjects learn causal relationships, reason and judge causally. Much of this work has been carried out by psychologists, but it also includes contributions by primatologists, researchers in animal learning, and neurobiologists, and more rarely, philosophers. In my view, the interactions between these two groups—the normative/theoretical and the empirical -- has been extremely fruitful. In what follows, I describe some ideas about causal reasoning that have emerged from this research, focusing on the "interventionist" ideas that have figured in my own work. I also attempt to extract some general morals regarding the kinds of interactions between the empirical and the more traditionally philosophical that in my experience have been most fruitful. Along the way (since this is a volume on experimental philosophy) I will also compare this research with some research strategies employed in X-Phi. I begin with this last topic (Section 2), turn then to some general themes about experimental methodology and what can be learned from experiment (Sections 3-4), and finally to more detailed illustrations (Sections 5- 7).

2. X-Phi, Survey Experiments, and ERC.

[*] Thanks to David Danks, Clark Glymour, and Edouard Machery, as well as three anonymous reviewers for OUP for extremely helpful comments on an earlier version of this essay.

In a broad sense, the phrase "experimental (or empirical) philosophy" might be taken to characterize any attempt to bring empirical results to bear on a philosophical issue. With this understanding, a great deal of work in philosophy of science (e.g., attempts to use special and general relativity to settle "philosophical" questions about the nature of space and time) and many projects in ethics and political philosophy pursued in a naturalistic spirit (e.g. Kitcher, 2011) might qualify as "experimental philosophy". However, this phrase is commonly used much more narrowly in contemporary discussion, to encompass experiments carried out by philosophers, often survey-like in character, with adult humans as subjects, and often organized around concerns about the role of "intuitions" in philosophical argument. Even conceived in this narrower way, X-Phi is (obviously) far from homogenous, but one useful distinction[1] contrasts (i) the use of experiments in a primarily negative or debunking role— to discredit appeals of more traditional philosophers to "intuition" or to claims about "what most people think" – and (ii) their use in the more positive role of providing evidence supporting claims about human psychology, including claims about the characteristics of various "concepts" people possess, patterns of reasoning and judgment in which they engage, and the psychological processes that generate these[2]. As several writers (e.g., Alexander, Mallon,

[1] I take this contrast between the negative and positive programs in X-Phi from Alexander, Mallon, and Weinberg, 2010, although, as they note, other writers employ a similar distinction. Examples of the positive program include the use of responses to vignettes to support claims about the folk concept of free-will (and whether or not it is compatibilist—e.g. Nichols and Knobe, 2007), and the folk concept of intentional action (e.g. Knobe, 2003). Examples of the negative program include Alexander and Weinberg, 2007 (challenging claims made by analytic epistemologists about the folk concept of knowledge) and Machery et al., 2004 (challenging the universality of Kripkean intutions about reference). The negative program also includes papers presenting evidence that the intuitive judgments of the folk or philosophers are influenced by such normatively irrelevant factors as order effects (e.g. Swain et al. 2008), context effects, and small variations in wording.

[2] I concede that some work in X-phi does not fall neatly into just one of these categories but still think that the distinction captures something epistemically important. Contrary to what some have claimed, as I see the negative project, it does *not* require commitment to the idea that X-phi results measure or detect anything like what armchair philosophers claim to be detecting through their "intuitions", or even that such intuitions (conceived, at a minimum, as requiring stable response patterns across and within subjects) exist or are sources of evidence about anything of philosophical interest. In particular, one can think of the negative project as resting on the claim that the intuitions of armchair philosophers are in the same epistemic position as the reports of X-phi subjects; if this is correct, the debunking project can succeed even if what is measured in X-phi surveys is, philosophically speaking, "noise" (or at least unstable etc. in a way that makes it unsuitable as evidence for traditional philosophical claims), since what this suggests is that the intuitions of the armchair philosopher are equally problematic. By contrast, the positive project takes on more of the commitments of traditional philosophical methodology, replacing the intuitions of single philosophers with the judgments of a larger number of folk, but retaining the idea that these are important sources of

and Weinberg, 2010) have noted, X-Phi results might be largely successful in their debunking role even if they have important limitations as sources of evidence for these more positive projects. On the other hand, the positive project is attractive for many reasons, not least because it is constructive and does not just consist in amassing negative results about claims of armchair philosophers. In what follows, I will largely focus on the use of X-phi in the service of this more positive project and how this relates to ERC.

I begin with some comparative remarks. Both X-Phi and ERC involve experiments in the minimal sense that variables of interest are manipulated and the results of such manipulations observed. As in X-Phi, ERC uses a number of subjects, rather than a single one or an unsystematically selected small group (the armchair philosopher and colleagues.) There are also, however, differences worth remarking. As observed above, a great deal of X-Phi consists in gathering verbal responses of adult humans (often college students) to hypothetical scenarios or vignettes, also described verbally. For example, subjects may be presented with a verbally described scenario and then asked whether they would judge that *C* caused *E* in the scenario. Such scenario-based experiments are common in some areas of psychology as well. Although I agree that such experiments can produce valuable results (see below for examples), it is also widely recognized (both in the X-Phi literature and by other commentators[3]) that they have a number of potential limitations when treated as sources of evidence regarding positive theses about human cognition: these include lack of experimental control over the way that subjects interpret the verbal scenarios and the likelihood that different subjects may interpret the scenarios differently, with these differences perhaps being correlated with other differences among subjects, possible sensitivity of subject responses to seemingly small variations in

evidence—e.g., about psychological phenomena or folk concepts of these. This is a source of concern because the problems with the traditional methodology go well beyond its *N=1* commitments. A closely related point is that while the focus of the negative project on whether the folk judge in connection with hypothetical examples as philosophers claim makes a great deal of dialectical sense, it is less obvious that the positive project is best pursued by looking at folk responses to such examples. For example, if one merely wants to challenge what philosophers claim about how people would judge in connection with true-temp (a scenario due to Lehrer, 1990 in which a device is implanted in S's brain in such a way that he reliably has beliefs that track the true temperature, without his knowing that or why he is reliable, the question being whether S "knows" the temperature), a survey may be perfectly sensible, since it may show, contrary to what Lehrer claimed, that there is no consensus among the folk that this is not a case of knowledge. However, surveying responses to true-temp may be less well-motivated if one's aim is to probe folk conceptions of knowledge, if, as I suspect is the case, the scenario involves factual assumptions that radically depart from those assumed when the folk concept applies. In general, when one engages in the positive project it is unclear why the questions asked and the sources of evidence considered should be restricted in the way that they often are in X-phi—that is to verbal responses to hypothetical scenarios and so on.

[3] See, e.g., Scholl, 2012.

wording[4], and order and other contextual or conversational effects (as when subjects attempt to guess what sort of response the experimenter is looking for and then provide it). In my view, the proper response to these limitations is not to eschew survey-style experiments, but rather to recognize that such experiments are most likely to be valuable when (i) combined with other sorts of evidence, including evidence about non-verbal behavior (see below), (ii) integrated with background knowledge from other sources, and (iii) addressed to appropriate questions, with due regard for the likely variable and contextual nature of the psychological phenomena one is trying to learn about. (Also below, section 4). Part of the appeal of some of the experimental work described/recommended below is that it scores well along these dimensions.

One particular way in which some ERC differs from typical X-phi survey experiments is that in the former, "scenarios" are, so to speak, realized materially rather than just described verbally. For example, subjects may be presented with various physical objects (toy airplanes, blickets and blicket detectors), rather than with just verbal descriptions of these. In some cases subjects may be asked to make verbal judgments about causal relations involving these objects but they may also be asked to engage in non-verbal behavior —for example, to "activate" the objects (e.g., Gopnik and Sobel, 2000).

Why does this matter? To begin with, there is substantial evidence that even adult humans sometimes respond very differently to stimuli that take the form of verbal descriptions of scenarios than to stimuli that are materially realized. As an illustration, subjects often perform differently and sometimes more optimally on statistical inference problems when presented with ecologically realistic data in the form of materially realized frequencies than when given word problems involving probabilities. Saffran et al. 1996 found that eight month-old children can learn to segment words from fluent speech just based on statistical relationships between neighboring speech sounds, even though intelligent adults are often remarkably bad at verbally presented reasoning problems involving probabilities. As a causal reasoning illustration, Danks et al. (forthcoming) found that although certain causal judgments about verbally described scenarios are influenced by information about norms, this effect largely disappears when subjects are asked to base causal judgments on materially realized frequencies.

A second point is that materially realized experimental designs often naturally allow for additional behavioral (and physiological/neural) measures besides the verbal

[4] For example, judgments about causation are known to be sensitive to the particular choice of words used in the verbal query employed by the experimenter -- see Collins and Shanks (2007) for examples. A closely related problem is that the same verbal query about causation may be interpreted differently by different subjects or may confound independent features the researcher is attempting to measure. For example, the common practice of asking subjects to rate how much they agree with various causal claims may confound assessments of causal strength with subject's degree of confidence that a causal relation of any sort is present and perhaps also with judgments about how good or paradigmatic example of causation the claim in question is, to the extent this is different from assessments of strength and confidence. If researchers are going to continue to use such verbal probes, they need to pay more attention to what the probes measure.

responses on which X-phi often focuses[5]. Most obviously, researchers can record whether subjects succeed or not at the experimental task, and this can tell us something about the structure and extent of their causal understanding, beyond what is suggested by their verbal behavior. Other measures of non-verbal behavior and processing can also be employed —e.g., reaction times, looking time measures, neuro-imaging techniques and so on. This helps to deal with some of the problems with exclusive reliance on survey results such as possible sensitivity to the particular choice of verbal description employed. Another important advantage of such "material" designs is that they allow for the use of subjects who cannot speak such as pre-verbal children and non-human animals. Including such subjects enables comparisons of causal cognition across species and also allows one to trace developmental trajectories of aspects of causal cognition among humans, as several of the experiments described below illustrate[6]. This greatly broadens the range of issues that can be explored[7]. Finally, materially realized experiments also naturally allow for a focus on normatively *successful* (or unsuccessful) performance, according to such fairly uncontroversial criteria for success as whether the subject succeeds in activating some device. This in turn provides information relevant to *learning,* which is important for reasons to which I now turn.

3. Methodological Reflections: The Role of Normative Theory, the Importance of Explaining Success, and the Function of Causal Thinking.

[5] Just to be clear, I am not claiming that the use of verbally described scenarios precludes the use of such other measures. For example, Joshua Greene's well-known experiments on moral reasoning measure, in addition to subject's verbal responses to hypothetical scenarios, both neural hemodynamic response via fMRI and reaction times. But (i) most experimental philosophers do not employ such alternative measures and (ii) when the task is one that calls for non-verbal as well as verbal behavior there are typically more quantities to measure—reaction times associated with the non-verbal behavior, various physiological measures associated with this behavior and so on.

[6] Two additional points: First, I am sensitive to the fact that many experimentally inclined philosophers will lack the training and resources to conduct non-survey experiments. The obvious solution is to collaborate with those who do have such training and resources. Second, it is of course true, as Edouard Machery has emphasized to me, that it may be technologically impossible or unethical to materially realize some scenarios—for example, one can't explore experimentally whether subjects will push people in front of real trolleys—and in such cases the use of hypothetical scenarios may seem unavoidable. Even in such cases, however, I think it is often worthwhile to try to experimentally create morally permissible real scenarios that are analogous to the hypothetical ones. For example, one can explore how subjects choose in "real" scenarios in which money or other material goods such as food (rather than trolleys) are diverted from set of people to another, as in Hsu et al., 2008.

[7] Of course, verbal responses to vignettes also sometimes can be evaluated according to whether they are correct of not according to some normative theory. My point is not that only "material" experiments allow for such evaluation (or that all material experiments automatically do), but rather that such experiments often can be designed in such a way that they provide uncontroversial benchmarks for successful performance

In my view, much of the most successful empirical, descriptive work on causal cognition has been guided by theories that are normative in the sense that they purport to describe how one ought to learn and reason about causal relationships, given epistemic goals like truth, predictive accuracy, and so on. By "guided" in this context I mean (among other things) that such theories have played an important role both in suggesting experiments and in the interpretation of experimental results. Examples of such work include Gopnik et al., 2004 and Griffiths and Tenenbaum, 2009 as well as other research cited below. By contrast, empirical investigations that are not seriously theory-guided at all or that seem motivated mainly by an interest in "refuting" theory-motivated claims seem to me to have produced less interesting (and often less easily interpretable) results[8]. Given the usual philosophical tendency to contrast normative and descriptive considerations and the idea that the aspirations of psychology and experimental philosophy are descriptive, such claims about the role of normative theory may seem surprising. In order to motivate them, I begin with the observation that human causal cognition seems to be fairly successful in enabling us to successfully get around in and cope with the world. To the extent this is so, there must be some story about *how* we are able to do this (that is, how we *succeed*) and this is something that it would seem can only be addressed by a combination of normative theorizing and descriptive results.

Here an analogy with "normative" or "ideal observer" theories of visual perception is suggestive[9]. The visual system is not just a set of mechanisms for producing "visual judgments" -- instead one of the most striking features of the visual system is that, although sometimes subject to visual illusions, it produces outputs that are largely reliable in the sense of enabling us to get around in the world successfully and providing accurate enough information about our local surroundings. Any adequate descriptive theory of the visual system needs to explain this fact – that is, to explain how, from the very limited information that impinges on the retina, the brain is able to reach conclusions that are veridical enough for many practical purposes about the three dimensional world of medium sized objects that lies around us. To explain how this is possible, investigators rely on theories that are normative (and typically computational) in character — "normative" in the sense of specifying computations and algorithms that show how it is possible to, e.g., derive accurate information about object segmentation from information available to the visual system concerning shading, edges and so on. To the extent that the visual system can be shown empirically to operate in accord with principles that we know are reliable in the sense of often issuing in accurate reconstruction of the visual scene before us, we have the basis for an explanation of why the visual system "succeeds". Obviously, to carry out this project we need, among other things, to think of the visual system and the principles by which it operates in broadly functional terms—the (or a) function or "goal" of the system's operation is to produce outputs that contain accurate enough information about aspects of the surrounding environment.

[8] Arguably, much of the psychological literature on causal attribution –e.g., Ahn et al. 1995 - illustrates this point. In my view the absence of a generally accepted normative account of actual causation has had a negative effect on work in this area.

[9] This analogy is due to Alison Gopnik.

Analogous projects have been pursued in connection with causal learning and causal reasoning. For example, normative theories of causal learning are available that purport to describe strategies, computations, algorithms, and background assumptions that enable one to reliably infer causal conclusions from various bodies of evidence, such as statistical information gained from passive observation and/ or information about the results of interventions. One may use such normative theories to explore the extent to which different subjects succeed or fail at various causal inference tasks and whether, when they succeed, they do so by making use of the sorts of strategies and patterns of reasoning described by the normative theory. Examples of this sort of work include Gopnik et al. (2001) and Sobel et al. (2004).

I contend it is often fruitful to investigate *causal judgment* in a broadly similar way. In this case the relevant normative framework will include a specification of the functions or goals of various forms of causal thinking and how distinctions that people make in causal judgment make sense in the light of these purposes. Empirical evidence about causal cognition is both interpreted in the light of this framework and used (in part) to evaluate different hypotheses about the considerations that guide people in making the causal judgments they do. This leads to a different focus and a different set of research questions from those that animate traditional metaphysical explorations or exercises in conceptual analysis, even for those who hold that empirical evidence is relevant to such matters. For example, rather than asking, as metaphysicians might, whether relations of double prevention are "really" causal, or asking, as a conceptual analyst might, whether it is part of "the concept" of causation that causation requires a connecting process, and then exploring whether empirical results might somehow bear on these matters, we instead focus on questions like the following: to the extent that people judge that some relations of double prevention are less than paradigmatically causal (in comparison with, e.g., cases of rocks shattering when hit by bottles), what point or goal might be served by this practice? Why, in terms of epistemic and other goals associated with causal judgment, do people care whether a connecting process is present or not when presented with a relation of counterfactual dependence? What factor or factors might people be tracking when they make distinctions among different relations of non-back-tracking counterfactual dependence, labeling some as causal and others as non-causal? Do subjects judge that all relations of double prevention are non-causal or only some and, if so, what factors seem to influence this distinction? In pursuing these sorts of functional inquiry, we needn't commit ourselves to the idea that either armchair philosophers or the folk (as revealed in surveys) possess "intuitions" that serve as a special source of information or insight about causation. We needn't even suppose that there is a single "concept" of causation that we are investigating, as opposed to a range of patterns of reasoning and judgments whose rationales we are attempting to understand (See Section 4).

I will add that interventionists about causation like me (Woodward, 2003) think the acquisition of information relevant to manipulation and control is among the goals centrally associated with causal thinking, and we thus look (at least in significant part) to these in order to understand empirical patterns in people's causal judgments[10]. I try to illustrate this idea below.

4. What Can We Expect to Learn: The Significance of Cognitive Variability and the Enabling Role of Theory.

Although some philosophers (e.g. Paul, 2010) suggest that empirical results from X-phi (and presumably ERC as well) can provide information about extra-mental phenomena—about causation or knowledge, "as they are in themselves", rather than merely how they are conceptualized by us, I will focus on the more widely accepted idea that insofar as there is a positive role for X-Phi results, it will consist in telling us something of philosophical interest about mental or psychological phenomena— how we think or reason about causation, free-will, reference, and so on. Assuming that this is the goal, what might we reasonably hope to learn from survey and other sorts of experimental results about causal cognition?

My first observation is that what we can expect to learn is constrained by the character of the phenomena we are investigating. Although not everyone agrees, I believe that empirical results from X-phi itself and many other considerations strongly suggest that there often is a great deal of variation across individuals in so-called higher cognition. Moreover, the performance of single individuals often shows strong contextual effects and intra-person variation in the sense that the same individual may make different judgments and use different cognitive strategies even in connection with what may look to researchers like closely related tasks. As illustrations of both points, there is good evidence that some significant proportion of experimental subjects (at least in our culture) reason in connection with some problems as though they are good Bayesians while other subjects use simpler, non-Bayesian strategies (e.g., reinforcement learning) in updating their beliefs[11]. In addition, the same subject may also reason correctly, by Bayesian standards, when presented with evidence in certain formats in certain contexts or in connection with certain problems and yet violate Bayesian rationality in other circumstances. For this reason, in my view it is usually not illuminating to ask such broad-brush questions as whether people are Bayesian or not—the correct view of "how Bayesian we are" will need to be far more nuanced. As another example, illustrated by the experiments of Walsh and Sloman and Lombrozo discussed in **7** below, we may find that a majority of subjects in one experiment judge that certain examples of double prevention (cases in which *c* prevents *d* which, had it occurred, would have prevented *e*) do not amount to causation, while a majority of subjects in a

[10] I acknowledge (how could I not?) that this normative, "functional" approach to understanding various forms of cognition can be misused, with the invention of ad-hoc just-so stories that attempt to rationalize this or that aspect of human thinking as serving some function where there is no independent evidence for this claim. This is a reason for being critical and tough-minded in evaluating such claims but not for eschewing them altogether.

[11] See, e.g., Grether and El-Gamal, 1995. The positive case that human subjects behave as Bayesians in causal learning tasks is made by Josh Tenenbaum in many papers—see, e.g. Griffiths and Tenenbaum, 2009.

different experiment judge that other examples involving double prevention *are* causal. This does not, in my opinion, show that ordinary people are confused or inconsistent or that the two groups of subjects operate with different concepts of causation. Rather it shows their thinking about double prevention is more complex and subtle than anything captured by simply asking whether "our" conception of causation is such that we regard double prevention relations as causal[12].

To the extent that the sort of variability and context-dependence just described is a feature of human cognition, it raises obvious questions about how we should think about the psychological phenomena researchers hope to detect in X-phi experiments (and in similar experiments in cognitive psychology). These questions are particularly pressing because, as already indicated, both in X-phi and related discussion, there is a tendency to express results as claims about "concepts" (as in, "the folk concept of free will is incompatibilist" or "our concept of causation requires the presence of a connecting process because cause and effect"), where the assumed notion of "concept" is what Wilson (2006) calls the "classical" notion. According to this notion, concepts are relatively stable, context-independent, widely shared informational structures with crisp and well-defined application conditions that are readily accessible to users. The central problem with this picture, as Wilson argues, is that a great deal of human cognition seems far more fluid, flexible and context-sensitive than what is described by the classical theory, with strategies and representational structures being more or less well-adapted to local cognitive niches but performing less well outside of these contexts, with the result that they are sometimes replaced with new strategies and structures in new contexts. Very often the successful deployment of a reasoning strategy or "concept" rests on a background edifice of empirical facts in a way that is very unobvious to users and that may only become apparent when they attempt to move outside the usual range of application of the strategy or concept. One consequence is that it may be difficult to settle, in a non-arbitrary way, what sorts of assumptions or commitments are truly constitutive of a concept or when one is dealing with a single concept that has been stretched or modified in an effort to adapt it to new circumstances and when instead we are dealing with several different concepts. This point has obvious implications for such issues as whether we operate with several different concepts of causation, as Hall (2004) has claimed, or only one [13].

[12] In a very interesting paper, Genone and Lombrozo (2012) show that with respect to concept reference, subjects are neither pure causal theorists or pure descriptive theorists but instead the same individuals rely on both descriptive and causal information to varying degrees in different contexts. My suggestion is that a similar pattern will be found for many other concepts of philosophical interest—reliance on "mixed" or "hybrid" theories, with considerable intra-individual variation depending on context.

[13] Another consequence is that verbal probes involving science fictionish scenarios (e. g., "true-temp" as a possible case of knowledge, alleged causal relationships involving magic) that violate background empirical assumptions for the application of a concept may produce results that tell us little about the structure of that concept. To the extent that people operate with non-classical concepts, requiring for their successful application, empirical assumptions of which users may be unaware, X-Phi results are more likely to tell us something about such concepts when non-outlandish scenarios are employed.

Wilson's claims about concepts and the empirical results about variation and context –dependence in cognition gestured at above seem to me to be mutually reinforcing. Taken together, they suggest it may not be fruitful to try to express "positive" results in X-phi (and related results in psychology) as claims about "our concept of X", at least if the notion of concept in play is anything like the classical one and the candidate Xs are at the level of generality of, e.g., "causation" or "free will". I would add that there are many interesting issues and problems connected to causal cognition that are not naturally viewed as having to do with our concept of causation in any sense, even if this is construed in a non-classical way—these include, for example, many issues about causal learning and causal reasoning strategies. They also include the role of various default assumptions. The latter may be very important in guiding causal reasoning in particular contexts even if the assumptions are not "constitutive" of any concept of cause shared by all or most of us[14].

The moral that I draw from these observations is *not* that we should give up on the empirical psychology of higher cognition or that any attempt to generalize in this area is misguided but rather that we need to reconceive the sorts of claims we can hope to establish. Applied to the subject of causal cognition, I suggest that we should be skeptical of (and not aim at either trying to establish or refute) sweeping generalities like " according to the folk concept of causation absences are causes" or "the folk have two distinct concepts of causation: dependence and production". Instead, I suggest that it is more fruitful to explore empirical claims of the following general sorts:

4.1a) Claims about what subjects of various kinds "can do" or "do with some frequency", particularly when it can be shown that other sorts of subjects rarely or never do this. For example, as discussed in section **5** below, many adult humans are able to distinguish between intervening and conditioning in normatively appropriate ways, learn causal relationships from their own and others interventions, and combine evidence from interventions and other sources in making causal judgments. However, by no means all adults do this, and other subjects (e.g., non-human primates or young children) may not be able do the latter at all. Results of this sort about what subjects can do are related to the ideas about the role of normative theory in explaining successful performance in 3 above. Again, we needn't express such claims as claims about the concepts possessed by most or all subjects—we can instead just talk in terms of abilities and reasoning patterns many subjects possess, how these conduce to success or failure given the goals they have and so on.

4.1b) Claims of the following sort: although subjects don't always do X (where X may be, e.g., judging or reasoning in a certain way), *when* they do X, this is how most or a substantial number of them do it. For example, by no means all subjects behave as Bayesians when faced with causal inference problems, but there is evidence that some

[14] For example, as Kushnir and Gopnik (2007) show, small children tend to assume as a default that many causal relations require spatial contact between cause and effect (or at least they learn such relations more readily) but they also readily give up this expectation when the evidence indicates, and learn causal relationships in which there is no spatial contact. If we simply ask whether it is part of the children's "concept" that causes must be in spatial contact with their effects, the answer is apparently "no", but this question misses the role the default assumption plays in the children's reasoning.

significant number do and that when they do, they are able to make accurate causal judgments by taking base rate information into account in the way prescribed by Bayes' theorem—see Sobel et al., 2004).

4.1c) Claims about factors affecting *variation* in judgment, reasoning, learning and so on[15]. In other words, rather than trying to establish quasi- universal claims like: the folk concept of causation has feature F (and perhaps treating apparent deviations from this concept as noise or error), one should instead focus on variations or differences in judgment etc. across different contexts and kinds of subjects and try to understand the factors affecting this variation.

4.2. The Enabling Role of Theory. Philosophers of science who study experimentation recognize many possible interactions between "theory" and "experiment". The use of experiments to "test" theories is just one possibility. I observed above that normative/philosophical theories of causal cognition can suggest features F such that if (as an empirical matter) we find them in people's reasoning, their presence can help to explain why that reasoning is successful, to the extent that it is. In a number of cases, it probably would not occur to anyone to do an experiment exploring whether these features are present in the absence of the normative theory[16]. This represents another respect in which philosophical or normative theory can play a positive or constructive role in experimental work: theory can *enable* or *motivate* or provide a rationale for doing certain experiments and a basis for interpreting their results. The results themselves may be interesting, surprising and valuable even for those who are skeptical of the motivating theory understood as a highly general hypothesis— it is in this sense that the role of the experiment is not just one of testing. This sort of construal will seem particularly natural if, as urged above, we think of many experimental results as more like demonstrations of what subjects can do or do with some frequency than as tests of hypotheses regarding what subjects always do— an illustration is provided by the results due to Bonawitz et al. below showing that 4 year olds are able to use correlations derived from passive observations to design novel interventions, while 2 year olds are not (or at least find this task far more difficult). Such results should seem interesting even to those who are skeptical of interventionism as a general theory of causation or causal judgment[17].

[15] As Machery has pointed out to me, a fair amount of work in X-phi takes this form, although more universalistic claims are also not uncommon.

[16] Of course I don't mean that in such cases it is *logically impossible* for anyone to think of the experiment in the absence of the motivating theory—merely that as a matter of empirical fact this usually does not happen.

[17] The converse possibility is also worth mentioning: Suppose we find experimentally that people's causal cognition, when successful, exhibits features G, where G is some feature that is not assigned a role in any current normative theory. This might motivate us to construct a new normative theory that explains the role of G in successful causal cognition. For what it is worth, the ideas about "causal specificity" developed in Woodward, 2010 had this sort of origin. I first noticed that biologists seemed to care about whether causal relations were specific or not without having any idea about whether there was any normative rationale for their doing so. This in turn suggested a search for such a rationale.

5. Interventionism

With this as background, I turn to a more detailed look at some empirical work on causation, beginning with some investigations that can be thought of as suggested or motivated by a broadly "interventionist" account of causation. According to this account, causal claims describe relationships that are potentially exploitable for purposes of manipulation and control. The version of this idea defended in Woodward, 2003 proposes the following connection between causation and intervention:

(**M**) Suppose *C* and *E* are variables. Then *C* causes *E* if and only if there is some intervention that changes the value of *C*, such that if that intervention were to occur, the value of *E* or the probability distribution of *E* would change.

This notion corresponds to what Woodward, 2003 calls a "total cause", which has to do with the total or overall effect of one variable on another[18]. An intervention on *C* with respect to a second variable *E* is an exogenous change in the value of *C* that is appropriately unconfounded from the point of view of inferring whether there is a causal connection from *C* to *E*. Roughly speaking, an intervention on C with respect to E changes *C* in such a way that any change in E, should it occur, occurs only through the change in *C* and not in some other way—any "route" by which *E* changes goes through *C*. (For details, see Woodward, 2003, pp. 98ff). Randomized experiments are one kind of intervention. As an empirical matter some human actions qualify as interventions but from the point of view of normative theory what makes a process count as an intervention is the causal structure of that process, whether or not it involves human action.

A crucial feature of this framework is that *intervening* on a variable is different from *conditioning* on it. In a common cause structure in which *C* is a common cause of E_1 and E_2, these last two variables will be correlated. However, intervening to "set" the value of E_2, via an intervention *I*, breaks the causal connection between *C* and E_2, with the result that E_1 and E_2, are no longer correlated under this intervention. Within the interventionist framework it is facts about how E_2 responds to an intervention on E_1 that track whether E_1 causes E_2.

(**M**) should not be understood as the obviously false claim that the only way to learn about causal relationships is by performing interventions. It is consistent with (**M**) that there are many different possible sources from which one can learn about causal

[18] This notion of total cause is meant to capture the idea of one variable having a non-null aggregate or overall impact on another variable over all the various routes or paths connecting to the two variables. A variable can have a causal influence on another along a particular path (and hence be what Woodward, 2003 calls a "contributing cause") even if it has no aggregate or total influence, as when influences along two paths cancel. The "only if" part of **M** holds for total causes but not for contributing causes (Woodward, 2003, p. 59.)

relationships, including learning from passive observation (not involving interventions). Construed as a normative claim, (**M**) implies that when one learns about a causal relationship (from whatever source), one should think of the content of what is learned in accord with (**M**). Moreover, our reasoning about causal relationships (from whatever source we learn about them) ought to be guided by the commitments implicit in (**M**).

My interest in what follows is not in defending (**M**) as a normative theory (for that see Woodward, 2003) but rather in exploring the extent to which, in addition to its normative credentials, (**M**) might figure in a descriptive empirical theory of how at least some subjects think and judge. For example, the close connections between causal claims and claims about what happens under interventions embodied in (**M)** naturally suggests empirical questions like the following:

> Can/do subjects learn about causal relationships in a way that respects the normative connection between causation and intervention embodied in (**M**)? That is, do subjects draw the causal conclusion suggested by (**M**) when given evidence that *Y* changes in value under an intervention on *X*?
>
> Are subjects sensitive to the differences between causal conclusions that are warranted when given evidence that *X* changes under an intervention and the conclusions warranted when *X* changes as result of some process that is "confounded" and hence not an intervention?
>
> Are subjects sensitive to the normative difference between intervening and conditioning? For example, are they more willing to infer that *X* causes *Y*, when given evidence that *Y* changes under interventions on *X*, than when given evidence that *X* and *Y* are correlated but which does not involve information about interventions?
>
> Do subjects treat their own actions as "default" interventions in the sense that when such actions produce a change in *X* that is associated with a change in *Y*, they tend to infer that *X* causes *Y*, in the absence of evidence that their actions are confounded in some way?

Questions of this sort can be explored empirically for many different sorts of subjects, including humans of various ages and non-human animals. There is evidence that the answers to all four questions is "yes" for adult human beings. (See, e.g., Steyvers et al, 2003, Lagnado and Sloman, 2004, Sloman and Lagnado, 2005.) Somewhat amazingly, at least to me, it also has been claimed on the basis of recent experiments that the answer to the third question is "yes" when the subjects are rats, although the interpretation of this experiment is controversial and there is no evidence that rats use co-variational information to design novel interventions in the way that humans are able to do. (Blaisdell et al., 2006)[19].

[19] Standard "associationist" models of learning, such as the Rescorla- Wagner model and its descendants are insensitive to the difference between intervening and conditioning in the sense that they treat information from both sources equivalently. Since human beings

These are some simple illustrations of experiments/ empirical investigations "suggested" or motivated by a philosophical, normative theory connecting causation and intervention. In the absence of such a normative theory, it is less likely that it would have occurred to anyone to do these experiments[20]. At the same time when we find, experimentally, that people behave in the ways described, this suggests that at least sometimes they are operating with ways of thinking about causation that are connected to intervention in the way that normative theory suggests. This in turn suggests the connection between causal judgment and our interest in manipulation and control is not just an arbitrary association made by some philosophical theory but connects to features actually present in people's causal judgments.

6. Toddlers as Agent Causal Learners

A more extended illustration of these themes is provided by some of the results in Bonawitz et al. (2010), but here a bit more background is required. Woodward (2007) observes that once one begins thinking about causation in interventionist terms, it is natural to distinguish conceptually among the following three possibilities:

are sensitive to this distinction (and it is normatively correct for them to be so) this is strong prima-facie evidence that the structure of human causal knowledge cannot be understood in purely associationist terms. It is an attractive feature of the interventionist framework that sensitivity to the differences among causal structures that imply the same independence and conditional independence relations but differ in what they predict would happen under different possible interventions can be regarded as an important diagnostic of whether subjects possess distinctively causal representations rather than mere representations of patterns of association.

[20] David Danks has objected that while the general idea that there is a connection between causation and manipulation (or action) may well have played a role in motivating these and other experiments, more specific ideas about the connection between causation and intervention in the sense of Woodward, 2003 or (presumably) of the sort found in Pearl, 2000 or Spirtes et al., 1993 were not required. To this I have several responses. First, the connection that I claim is not that possession of a motivating theory was required as a matter of logic for anyone to do the experiments but rather that possession of the theory as a matter of causal empirical fact prompted or motivated people to do the experiments. And while it is certainly true that the general idea of a connection between causation and manipulation has been around for a long time, I don't think that the impact of well-worked out and in context novel computational ideas of the sort found in Spirtes et al. and in Pearl should be underestimated—these helped to turn the vague (and in philosophical circles much criticized) idea that there was such a causation/manipulation connection into something that looked conceptually and mathematically respectable. In my opinion, this in turn influenced experimental practice. Finally, there certainly are experiments that draw on specific ideas about what an intervention involves—for example, experimental demonstrations that subjects are sensitive to whether their manipulations are confounded in making causal inferences, as in Kushnir and Gopnik, 2005.

1) *Egocentric causal learning*: A subject might be able to learn about causal relationships from her own interventions but not from other sources. This might happen either because (i) the subject is not able to learn causal relationships at all from observing the results of other agent's interventions or from observation of contingencies not involving interventions or because (ii) she may learn such relationships but fail to integrate them into a single unified representation, not recognizing, e.g., that the results of the interventions of others provide evidence for what would happen if she were to perform similar interventions.

2) *Agent causal learning*: the subject grasps that the same relationship she exploits in intervening also can be present when other agents act. Accordingly, the subject learns causal relationships both from observing the results of her own interventions and from observing the results of the interventions of other agents, recognizing, for example, that the results of the interventions of others provide evidence for what would happen if she were to perform similar interventions. But the agent does not similarly learn causal relationships from "passive" observations of contingencies not involving interventions or fails to integrate these with what is learned from observing interventions.

3) Fully (or more nearly fully) causal learning: the subject grasps that the same relationship she exploits in intervening also can be present both when other agents intervene and in nature even when no other agents are involved and integrates information from each of these sources into a common unified representation. One empirical test for the presence of this ability is whether a subject will perform appropriate novel interventions to obtain some goal on the basis of an observed association between events not involving interventions—to use Tomasello's and Call's (1997) example, would an ape, observing the wind shake fruit lose from a tree, be able to infer that if it were to shake the branch, this would cause fruit to fall to the ground?

Adult humans are plainly full causal learners. An interesting empirical question, suggested by the distinctions above, is which of the three categories mostly closely captures the causal cognition of non-human animals or young human children. Some primatologists, including Tomasello have, in effect claimed that non-human primates are not full causal learners -- a claim which, if correct, would mark a fundamental difference between human and non-human causal cognition. But what about very young humans—babies and toddlers? Is there a stage in human development in which young children behave as egocentric or agent but not fully causal learners? Bonawitz et al. 2010 reports experimental results addressing this question. Compressing greatly, we compared children in two different age groups (toddlers vs. pre-schoolers, mean ages in first experiment 47. 2 vs 24.4 months) who observed a block slide toward a toy airplane which activated when the block touched its base. In the "ghost condition" the block appeared to move spontaneously, with no agent involved, activating the plane on touching. In another condition (the agent condition), an experimenter moved the block, demonstrating it would activate the plane on touching.

The children were then asked to make the airplane go by themselves. A very large majority of pre-schoolers were able to use their previous observations of the association between the movement of the block and the activation of the plane in the ghost condition to intervene on their own to activate the plane. By contrast, none of the toddlers was able to do this, although they were able to successfully activate the plane via their own interventions when they observed another agent do this.

We interpreted this and other experimental results as prima-facie evidence that there is a stage in the development of human causal cognition in which children (in this case toddlers) are able to learn to design their own interventions from observations of the interventions of other agents, but not from the observation of otherwise similar associations not involving interventions. In other words, at this point the children appear to be in something like an agent causal stage. This is then followed by a stage in which fully causal learning is achieved, as evidenced by the behavior of the older children. This interpretation is consistent with a great deal of other developmental evidence showing that even very young children seem to be very sensitive to the difference between agents and non-agents in their environments and primed for or adept at learning from the behavior of other agents (particularly the intentional or goal-directed actions of other agents) as opposed to other sources of information. At least one member of our collaboration (Meltzoff, 2007) thinks that as soon as children represent their own actions they also represent the actions of others in a common amodal code and hence in effect they are never in a purely egocentric stage of causal learning.

We may think of this as yet another example of an experiment motivated or suggested by a normative theory: unless one has the "philosophical/ computational" idea that there is an important connection between causation and intervention and that some human actions are paradigmatic interventions, one is unlikely to undertake the sort of investigation just described. Given this idea, and the accompanying thought that one sometimes learns about causal relationships from interventions, it is natural to ask how and when in development information from interventions is integrated with other sources into more unified causal representations.

This experiment may also be used to illustrate how influence can flow in the opposite direction—how empirical results can have implications for philosophical theorizing. Consider an "agency" theory of causation like that defended by Menzies and Price (1993), according to which our concept of causation is derived from our subjective experience of agency, which is then "projected" onto the world. When this is interpreted as an empirical claim, experimental results of the sort described above suggest there is something broadly right about the idea, in the sense that our capacities to act as agents and to manipulate plays an important role in the development of our capacities for causal learning and cognition. On the other hand, these experimental results also suggest the need for a somewhat more complex story than the one told by Menzies and Price. To see this, recall a standard philosophical criticism of agency theories: they face problems explaining how the notion of causation comes to be extended to unmanipulable causes, such as earthquakes where no one has the relevant experience of agency – a criticism which Menzies and Price address. However, the experimental results above suggest that a related problem arises, so to speak, much earlier—there is already a problem about how causal learner X moves from her own subjective experience of agency to a recognition that causal relations are present when other agents act and produce effects, since in such cases X presumably does not have the experience of agency. Moreover, there is a parallel problem about cases in which the subject infers the existence of a causal relationship involving causes that are straightforwardly manipulable, but which are not in fact manipulated by any agent, as when wind shakes fruit from a tree. "Projection" can function as a label for our willingness to extend the notion of causation from our own manipulations to such cases but it provides no insight into how we develop the capacity to do this or the factors influencing such development. Moreover, if Meltzoff is correct, it may be an empirical mistake to suppose that young humans begin in a purely "egocentric" stage centered around their own experience of agency—instead we may be agent-causal as soon as we are capable of any kind of causal learning. If so, although agency may be crucial to the development of causal cognition, it may be that what is central is not so much the *experience o*f agency, but rather other features associated with the capacity to act effectively as an agent such as the ability to represent means/ends structures of actions exhibited both by self and others.

7. Empirical Results Concerning Double Prevention

I turn next to a quite different topic— some recent experimental work on double prevention. These are cases in which if *d* were to occur, it would prevent the occurrence of *e* (which would otherwise occur in the absence of *d*) and in which the occurrence of *c* prevents the occurrence of *d*, with the upshot that *e* occurs. In Ned Hall's well-known example (2004), Suzy's plane will bomb a target (*e*) if she is not shot down by an enemy pilot (*d*). Billy, piloting another plane, shoots down the enemy (*c*), and Suzy bombs the target.

In such cases there is overall counterfactual dependence (of a non-backtracking sort) of *e* on *c*. This is taken to be sufficient for causation by many counterfactual theories, as well as by **M**. Nonetheless, many philosophers (and non-philosophers) find it intuitive that cases of double prevention either do not involve causation at all, or at least lack some feature which is central to some other cases of causation—e.g., the presence of a connecting process or transmission of energy/momentum. (Following Hall 2004, I will describe these as involving "production"). By contrast, other philosophers, such as Schaffer, 2000 emphasize that many biological mechanisms (from muscle contraction to gene regulation) and many manufactured artifacts (e.g., many guns) operate by relations of double prevention and that these relations can look paradigmatically causal.

This raises several questions. First, one might wonder, as a matter of descriptive psychology, to what extent, if any, people regard relations of double prevention as causal. Is it possible that some cases of double prevention are regarded as more paradigmatically causal than others? Second, to the extent that people distinguish between production and "mere" counterfactual dependence, why do the do so? What point or purpose is served by this practice?

Walsh and Sloman, 2011 presented subjects with a series of scenarios. In one, a coin stands unstably on edge, about to fall tails. Billy and Suzy roll marbles in such a way that, if nothing interferes, each will strike the coin so that it lands heads. Billy's marble strikes first. In this scenario, 74% of subjects judged that Billy's marble caused the coin to land heads.

In a second scenario, an unstable coin is again on edge, about to land heads. A third party rolls a marble in such a way that if it strikes the coin, it will land tails. However, a book blocks the path of the marble. Frank removes the book but if he had not, Jane would have. The marble strikes the coin and it lands tails. In this scenario only a minority of subjects (38%) judge that Frank caused the coin to land tails and virtually no one judges that Jane caused this effect[21].

Walsh and Sloman's Billy/Suzy scenario is a case of causal preemption in which the effect is not counterfactually dependent in any simple, straightforward way on the cause but in which transmission or a connecting process is present between Billy's throw and the coin's falling heads. In the second scenario, the relation of Frank's action to the coin's falling tails is double prevention-like, with the added complication that because Jane's action is a pre-empted double preventer, the coin's falling tails is not counterfactually dependent on Frank's action. Although Walsh and Sloman do not report any results about experiments with a double prevention structure without a second, pre-empted preventer (i.e., the same scenario as above with Jane absent), it is a reasonable guess that fewer subjects would judge that Frank's action was a cause in such cases than in a variation on the first scenario in which only Billy is present. These results thus seem to support the conclusion that as far as folk thinking about causation goes, cases in which a "connecting process" is present are more likely to be judged as causal than cases of double prevention. Indeed, this is one of the conclusions Walsh and Sloman draw. I will return to this claim below but I want first to describe some additional experiments, due to Lombrozo , 2010, that complicate matters in an interesting way.

Lombrozo's experiments explore people's causal judgments about double prevention scenarios in contexts involving intentional action, artifacts with designed functions, and biological adaptations. Although Lombrozo finds, in agreement with Walsh and Sloman, that subjects are more willing to describe cases involving production as causal than (at least some) cases involving double prevention, she also finds that subjects distinguish among cases of double prevention, treating cases involving intentional action, designed function and biological adaptation as more causal than cases of double prevention not involving these features. For example, she presents subjects with a fictitious example in which the tendency of a species of shrimp to reflect UV light depends, via a double prevention structure, on their diet. Subjects are more willing to judge that this double prevention structure is causal when they are told that this light-reflecting tendency is a biological adaptation than when it has no adaptive significance and similarly for parallel examples involving effects that are intended or the result of a designed function. Since none of Lombrozo's cases of double prevention involve an uninterrupted connecting process, or transfer of energy/momentum from the putative cause to its effect, it cannot be the presence or absence of these features that accounts for subjects' willingness to regard some cases as more paradigmatically causal than others.

Following a suggestion in Woodward, 2006, Lombrozo proposes that relations of double prevention are judged as more paradigmatically causal to the extent that they are more *stable.* Suppose that *E* counterfactually depends on *C*, where the dependence is

[21] Although less than half, 38% is a non-trivial number. How should we think about such subjects? Are they making a mistake, and misapplying "our concept" of causation, the content of which can be determined by the majority who did not regard this relationship as causal? Are they perhaps operating with a different "concept" of causation than the majority? The unsatisfactory character of either of these options illustrates the appeal to the alternative approach recommended above, in which one replaces these questions with such questions as why subjects sometimes distinguish between double prevention and other sorts of causal relations.

non-back- tracking and intervention- supporting in the manner described in **M**. Stability has to do with the extent to which this relation of counterfactual dependence would continue to hold as other factors in the background, in addition to *C* and *E,* change[22]. Both theoretical considerations (see below) and casual observation suggest that, other things being equal, dependence relations that are more stable are, as an empirical matter, judged as more paradigmatically causal than less stable relations. This effect is present both for examples that do not involve double prevention and for examples that do[23], and provides a natural explanation of Lombrozo's results. In particular, as argued both by Lombrozo and in Woodward, 2006, it is plausible that cases of double prevention involving biological adaptations, designed functions and intentional actions tend to be more stable under relevant changes than cases of double prevention lacking these features. For example, double prevention relations involving biological adaptations tend to be more buffered against environmental disruptions (and more stable for that reason) than double prevention relations that arise, so to speak, fortuitously. Somewhat more ambitiously, one might attempt to tell a similar story, also based on stability, about why relationships of counterfactual dependence involving transmission or connecting processes are commonly judged more paradigmatically causal than those lacking this feature.

But why should stability matter in the way described to judgments of causal status? We can provide at least a partial answer to this question by making use of the strategy recommended in section 3 and asking how a concern with stability fits in with goals or functions of causal thinking. According to interventionist accounts, one reason why people care about the difference between causal relationships and purely correlational relationships is that the former are exploitable (at least in principle and often in fact) for purposes of manipulation/intervention and control in a way that the latter are

[22] For reasons of space, I refer the reader to Woodward, 2006 for a more detailed discussion of stability. However, some brief clarificatory remarks may be helpful. First, stability comes in degrees and is relative to particular sets of background conditions. Stability claims thus have the form: relation of counterfactual dependence *C* is stable under such and such changes in background conditions *B* (but perhaps not under other changes in *B* or under changes in other background conditions *B'*). The idea is thus *not* to dichotomize claims as stable or unstable in some absolute sense or to pick out some single set of background conditions as uniquely appropriate in assessing stability. The account recognizes that virtually all relations of dependence will hold under some circumstances and not under others. Nonetheless, both as a normative matter and as description of how people think, some sets of changes are more important than others for purposes of assessing stability-- which sets depends both on a number of considerations having to do both with subject matter and which changes are, as a statistical matter, common or usual. In the examples involving biological adaptiveness discussed below, for example, biologically normal changes in the environment or in systems within the organism will be particularly important for judgments of stability.

[23] Cf. Lewis, 1986 who compares the unstable (or, as he calls it, "sensitive") causal relation between writing a letter of recommendation and the existence of certain people in the distant future with the much more stable relation between shooting at close range and death.

not. Identifying relationships relevant to intervention is, as Lombrozo says, one "function" or normative aim of causal ascription, even if it may not be the only such function. Once this idea is accepted, it is also natural to recognize that different relationships of counterfactual dependence (of the non-backtracking sort) can serve this function or satisfy this aim to different degrees. In particular, the more stable a relationship of counterfactual dependence is, then (other things being equal) the more useful or suitable it is likely to be for purposes of manipulation and control, the more exportable or generalizable it is likely to be to other contexts, and the more it is likely to continue to hold in the present context in the face of changing contingencies. Thus, at least within an interventionist framework, it "makes sense" and is normatively appropriate for subjects to distinguish in the way that they do among relationships of counterfactual dependence in terms of their degree of stability.

An additional attraction of this way of looking at things is that it allows us to move beyond the "dull thud of conflicting intuitions" that characterizes much philosophical discussion of double prevention and to ask some more productive questions. Rather than trying to appeal to "intuition" to settle whether double prevention relations are truly causal, we can investigate empirically the judgments people make in double prevention cases, the factors that influence these judgments, and we can also ask whether these judgments have some recognizable normative rationale. This allows us to understand why (or to what extent it might be rational for) people to operate with notions of causation that incorporate these features. In other words rather than arguing about whether causation (or our concept of causation) requires a connecting process, we ask what goals or purposes might be served by distinguishing among relations of dependence according to whether a connecting process is present, whether there might be some point to distinguishing among relations of dependence that lack such a process and so on. This seems to me to provide one way in which empirical results can be relevant to understanding causal reasoning that does not require problematic assumptions about the role of intuition or, for that matter, strong assumptions about the character of "our concept" of causation.

8. Conclusion

No deep sociological insight is required to recognize that one factor animating current disputes about X-Phi is anxiety about the status and future of philosophy itself. Put crudely, the worry is that philosophy is going to be replaced by something more like experimental psychology or at least that this is the future for philosophy advocated by some experimental philosophers. At least in the case of causation, this worry strikes me as misplaced; as I have tried to show, there remains lots (that is positive and constructive) for philosophers to do, even if philosophical work on causation becomes more influenced by empirical considerations. In particular, if the overall argument of this paper is correct, one reason why purely descriptive reports of what the folk or scientists think about causation are not going to replace normative, philosophical theorizing is that we need the normative theorizing to do the descriptive work properly.

More generally, although I have been skeptical of the value of asking broad brush questions about "our concept of cause" and similar matters, I believe philosophers have accumulated a battery of distinctions and an understanding of connections among different features of causal thinking that can play a very useful role in experimental design and interpretation. However obvious these distinctions/connections may seem to philosophers, they are frequently not well understood by experimental psychologists and other non-philosophers and this can lead to badly designed experiments and difficult to interpret results. As a simple illustration, not discussed in any detail above, consider the contrast most philosophers would draw between, on the one hand, (i) the project of capturing a broad, egalitarian notion of what it is for one factor to be causally relevant to another (where this contrasts with the factor being causally irrelevant, although perhaps correlated) and (ii) projects associated with causal selection, where this has to do with understanding the basis on which we select, from among those factors satisfying the broad notion under (i), one or more as, e.g., "a cause", "the cause" etc. of some effect. The target of many philosophical theories of causation, as well as most of the normative/computational theories of causal inference outside of philosophy, is (i) rather than (ii), while (ii) seems to be the typical target of investigations of causal attribution by psychologists. Whatever the details of one's views about (i) and (ii) it seems uncontroversial that these are *different* projects. If so, we should be wary of using empirical results about the considerations influencing causal selection to reach conclusions about the descriptive (much less normative) adequacy of claims having to do with (i). For example, even if it is true that causal selection is not guided by considerations having to do with counterfactual dependence, as suggested in Mandel, 2003, it does not follow that counterfactual theories of causation understood as contributions to project (i), interpreted either descriptively or normatively, are thereby refuted. More generally, the philosophical distinction between (i) and (ii) should allow us to see that it should not simply be assumed that results about causal selection and actual cause judgment (and the learning and processing that underlie such judgments) automatically generalize to other sorts of causal claims. Other, similar illustrations abound.

References

Ahn, W., Kalish, C., Medin, D., and Gelman, S. (1995) "The Role of Covariation versus Mechanism information in causal attribution" *Cognition* 54:299-352.

Alexander, J. & Weinberg, J. (2006). "Analytic Epistemology and Experimental Philosophy". *Philosophy Compass*.

Alexander, J. and Mallon, R. (2010) "Accentuate the Negative" *Review of Philosophy and Psychology* 1: 297-314.

Bonawitz, E., Ferranti, D., Saxe, R., Gopnik, A. Meltzoff, A., Woodward, J. and Schulz, L. 2010. "Just do it? Investigating the Gap Between Prediction and Action in Toddlers' Causal Inferences." *Cognition* 115: 104–117.

Collins, D. and Shanks, D. (2006) "Conformity to the Power PC Theory of Causal Induction Depends on the Type of Probe Question" *The Quarterly Journal of*

Experimental Psychology: 59 225–232.

Danks, D., Rose, D. and Machery, E. (forthcoming) “Demoralizing Causation”

Genone, J. and Lombrozo, T. (2012) “Concept possession, Experimental Semantics, and Hybrid Theories of Reference” *Philosophical Psychology*, 25: 717-742.

Gopnik, A., Glymour, C. Sobel, D., Schulz, L. Kushnir, T and Danks, D. (2004) “A Theory of Causal Learning in Children: Causal Maps and Bayes Nets”. *Psychological Review* 111: 3-32.

Gopnik, A, Sobel D., Schulz, L. and Glymour, C. (2001) “Causal Learning Mechanisms in Very Young Children: Two, Three and Four-year-olds infer Causal Relations from Patterns of Variation and Covariation” *Developmental Psychology* 37: 620-6.

Gopnik, A. and Schulz, L. 2007. *Causal Learning: Psychology, Philosophy and Computation*. New York: Oxford.

Kitcher, P. (2011) *The Ethical Project*. Cambridge: Harvard University Press.

Kusnir, T. and Gopnik, A. (2007) “Conditional Probability versus Spatial Contingency in Causal Learning: Preschoolers Use New Contingency Evidence to Overcome Prior Spatial Assumptions” *Developmental Psychology* 43: 186-196.

Grether, D. and El-Gamal, M. (1995) “Are people Bayesian? Uncovering Behavioral Strategies” *Journal of the American Statistical Association* 90:1127--1145.

Griffiths, T. and Tenenbaum, J. 2009. “Theory-Based Causal Induction.” *Psychological Review* 116: 661-716.

Hall, N. (2004) “Two Concepts of Causation” in *Causation and Counterfactuals* (ed. Collins, Hall, and Paul) Cambridge: MIT Press.

Hsu, M. , Anen, C. and Quartz, S. (2008) “The Right and the Good: Distributive Justice and Neural Encoding of Equity and Efficiency” *Science* 320: 1092-1095

Knobe, J. (2003). “Intentional Action in Folk Psychology: An Experimental Investigation”. *Philosophical Psychology, 16*(2), 309-323.

Kushnir, T. and Gopnik, G. 2005. “Young Children Infer Causal Strength from Probabilities and Interventions. *Psychological Science* 16: 678-683.

Lagnado, D. and Sloman, S.A. (2004) “The Advantage of Timely Intervention”. *Journal of Experimental Psychology: Learning, Memory & Cognition* 30: 856-876.

Lehrer, K. (1990) *Theory of Knowledge* Boulder, CO: Westview Press.

Lewis, D. (1986) “Postscripts to ‘Causation’”. *Philosophical Papers*, Vol 2. Oxford: Oxford University Press.

Lombrozo, T. (2010). “Causal-explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions”. *Cognitive Psychology* 61: 303-332.

Machery, E., Mallon, R. Nichols, S. and Stich, S. (2004) “Semantics, Cross-Cultural Style.” *Cognition* 92 :B1–B12.

Mandel, D. (2003) “Judgment Dissociation Theory: An Analysis of Differences in Causal, Counterfactual, and Covariational Reasoning”. *Journal of Experimental Psychology*: *General*, 132, 419-34.

Meltzoff, A. (2007) “ Infants’ causal learning: Intervention, observation, imitation”. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, Philosophy, and Computation.* Oxford: Oxford University Press, pp 37-47.

Menzies, P. and Price, H. (1993) “Causation as a Secondary Quality” *British Journal for the Philosophy of Science* 44: 187-203.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner. 2005. “Surveying freedom: folk intuitions about free will and moral responsibility”. *Philosophical Psychology* 18: 561–584.

Paul, L. (2010) “A New Role for Experimental Work in Metaphysics” *Review of Philosophy and Psychology* 1: 461-476.

Pearl, J. (2000) *Causality*. Cambridge: Cambridge University Press.

Saffran, J., Aslan, R. and Newport, E. (1996) “Statistical Learning by 8 Month-old Infants” *Science* 274: 1926-28.

Schaffer, J. (2000) “Causation by Disconnection” *Philosophy of Science* 67: 285-300.

Scholl, B. “Two Kinds of Experimental Philosophy (and their methodological dangers)” http//experimentalphilosophy.typepad.com/experimental_philosophy/files/scholl_xphi_notes.pdf. Accessed 8/11/12.

Sloman, S. and Lagnado, D. (2005). “Do we ‘do’?” *Cognitive Science* 29: 5–39.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). “Children's Causal Inferences from Indirect Evidence: Backwards Blocking and Bayesian Reasoning in Preschoolers.” *Cognitive Science 28*: 303-333.

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

Steyvers, M., Tenenbaum, J., Wagenmakers, E., and Blum, B. (2003) "Inferring Causal Networks from Observations and Interventions." *Cognitive Science* 27: 453-489.

Swain, S. , Alexander, J. and Weinberg, J. (2008) "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp" *Philosophy and Phenomenological Research* 56: 138-155.

Tomasello, M. and Call, J. 1997. *Primate Cognition*. New York: Oxford University Press.

Walsh, C. and Sloman, S. (2011). "The Meaning of Cause and Prevent: The Role of Causal Mechanism." *Mind and Language,* 26: 21–52.

Wilson, M. (2006) *Wandering Significance: An Essay on Conceptual Behavior* Oxford: Oxford University Press.

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

Woodward, J. (2006) "Sensitive and Insensitive Causation." *Philosophical Review* 115: 1-50.

Woodward, J. (2007) "Interventionist Theories of Causation in Psychological Perspective." In A. Gopnik and L. Schulz (eds.) *Causal Learning: Psychology, Philosophy and Computation.* New York: Oxford University Press, 19-36.

Woodward, J. (2010) "Causation in biology: stability, specificity, and the choice of levels of explanation" *Biology and Philosophy* 25:287–318.