Forthcoming in A. Gopnik and L. Schulz (eds.) *Causal Learning: Psychology, Philosophy and Computation.* New York: Oxford University Press.

Interventionist Theories of Causation in Psychological Perspective

Jim Woodward

1. Introduction. Broadly speaking, recent philosophical accounts of causation may be grouped into two main approaches: *difference-making* and *causal process* theories. The former rely on the guiding idea that causes must make a difference to their effects, in comparison with some appropriately chosen alternative. Difference-making can be explicated in a variety of ways. *Probabilistic* theories attempt to do this in terms of inequalities among conditional probabilities: a cause must raise or at least change the probability of its effect, conditional on some suitable set of background conditions. When probabilistic theories attempt to define causation in terms of conditional probabilities, they have obvious affinities with associative theories of causal learning and with the use of contingency information (conditional Δp) as a measure of causal strength (Dickinson and Shanks, 1995). *Counterfactual* theories explicate difference making in terms of counterfactuals: a simple version might hold that C causes E if and only if it is true both that: (i) if C were to occur, E would occur and (ii) if C were not to occur, E would not occur. Following David Lewis, counterfactuals are often understood in the philosophical literature in terms of relationships among possible worlds: very roughly, a counterfactual like (i) is true if and only if there is a possible world in which C and E hold that is "closer" or "more similar" to the actual world than any possible world in which C holds and E does not hold. A set of criteria is then specified for assessing similarity among possible worlds. (cf. Lewis, 1986, p.47)

The interventionist theory described in Section 2 is a version of a counterfactual theory, with the counterfactuals in question describing what would happen to E under interventions (idealized manipulations of) on C. The interventionist theory does not require (although it permits) thinking of counterfactuals in terms of possible worlds and, as noted below, the specification of what sorts of changes count as interventions plays the same role as the similarity metric in Lewis' theory. When causal information is represented by directed graphs as in Bayes net representations, these may be given an interventionist interpretation (Woodward, 2003, Gopnik and Shulz, 2004).

It is usual in the philosophical literature to contrast so-called type causal claims which relate one type of event or factor to another ("Aspirin causes headache relief") with token or singular causal claims which relate particular events ("Jones' taking aspirin on a particular occasion caused his headache to subside"). There are versions of difference-making accounts for both kinds of claim, although it is arguable that such accounts apply most straightforwardly to type causal claims. In contrast, causal process accounts apply primarily to singular causal claims. The key idea is that some particular event c causes some other event e if and only if there is a connecting casual process from c to e (Salmon, 1994). Pro-

cesses in which one billiard ball collides with another and causes it to move are paradigmatic. There are a number of different accounts of what constitutes a causal process, but it is perhaps fair to say is that the generic idea is that of a spatio-temporally continuous process that transmits a conserved quantity such as energy and momentum, or, as it sometimes described in the psychological literature, "force". Theorists in this tradition often deny that there is any intimate connection between causation and difference making: they claim that whether c causes e depends only on whether there is a causal process connecting c and e, something that (it is claimed) does not depend in any way on a comparison with what happens or would happen in some other, contrasting situation(s) (Salmon, 1994, Bogen, 2004). In contrast, such comparisons are at the heart of difference making accounts. Although most philosophical versions of causal process accounts are not committed to claims about the possibility of perceiving causal connections, an obvious analogue in the psychological literature are approaches that focus on launching or Michotte type phenomena. Psychological theories that attempt to understand causation in terms of mechanisms or generative transmission (where these notions are not understood along difference-making lines) are also in broadly the same tradition.

2. Interventionism. Interventionist accounts take as their point of departure the idea that causes are potentially means for manipulating their effects: if it is possible to manipulate a cause in the right way, there would be an associated change in its effect. Conversely, if under some appropriately characterized manipulation of one factor, there is an associated change in another, the first causes the second.

This idea has a number of attractive features. First, it provides a natural account of the difference between causal and merely correlational claims. The claim that X is correlated with Y does not imply that manipulating X is a way of changing Y, while the claim that X causes Y does have this implication. And given the strong interest that humans and other animals have in finding ways to manipulate the world around them, there is no mystery about why they should care about the difference between causal and correlational relationships. Second, a manipulationist account of causation fits very naturally with the way such claims are understood and tested in many areas of biology and the social and behavioral sciences and with a substantial methodological tradition in statistics, econometrics and experimental design, which connects causal claims to claims about the outcomes of hypothetical experiments.

Although it is possible to provide a treatment of token causation with a manipulability framework, I will focus on the general notion of one type of factor being causally relevant (either positively or negatively) to another. There are two more specific causal concepts that may be seen as prescifications of this more general notion: total causation and direct causation. X is a *total cause* of Y if and only if it has a non-null total effect on Y -- that is, if and only if there is some intervention on X alone (and no other variables) such that for some values of other variables besides X, there will be a change in the value of

Y under this intervention. Woodward, 2003 argues that this notion is captured by the conjunction of two principles (**TC**):

> (**SC**) If (i) there are possible interventions (ideal manipulations) that change the value of X such that (ii) if such an intervention (and no others) were to occur X and Y would be correlated, then X causes Y.
>
> (**NC**) If X causes Y then (i) there are possible interventions that change the value of X such that (ii) if such interventions (and no other interventions) were to occur, X and Y would be correlated.

Before turning to the notion of direct causation, several clarificatory comments are in order. First, note that if **TC** is to be even prima–facie plausible, we need to impose restrictions on the sorts of changes in X count as interventions or ideal manipulations. Consider a system in which A = atmospheric pressure is a common cause of the reading B of a barometer and a variable S corresponding to the occurrence/non-occurrence of a storm, but in which B does not cause S or vice-versa. If we manipulate the value of B by manipulating the value of A, then the value of S will change even though, in contradiction to (**SC**), B does not cause S. Intuitively, an experiment in which B is manipulated in this way is a badly designed experiment for the purposes of determining whether B causes S. We need to formulate conditions that restrict the allowable ways of changing B so as to rule out possibilities of this sort.

There are a number of slightly different characterizations of the notion of an intervention in the literature – these include Spirtes, Glymour, and Scheines, 2000, Pearl, 2000, and Woodward. 2003. Since the difference between these formulations will not be important for what follows I will focus on the core idea. This is that an intervention I on X with respect to Y causes a change in X which is of such a character that any change in Y (should it occur) can only come about through the change in X and not in some other way. In other words, we want to rule out the possibility that the intervention on X (or anything that causes the intervention) affects Y via a causal route that does not go through X, as happens, for example, when B in the example above is manipulated by changing the common cause, A, of B and S. I will also assume in what follows that the effect of an intervention on X is that X comes entirely under the control of the intervention variable and that other variables that previously were causally relevant to X no longer influence it – that, as it is commonly put, an intervention on X, "breaks" the causal arrows previously directed into X. In the case of the A-B-S system an intervention having these features might be operationally realized by, for example, employing a randomizing device which is causally independent of A and B and then, depending on the output of this device, experimentally imposing (or "setting") B to some particular value. Under any such intervention, the value of S will no longer be correlated with the value of B and (**NC**) will judge, correctly, that B does not cause S. Note that in this case, merely observing the values of B and S that are generated by the ABS structure, without any intervention is a very different matter from intervening

on B in this structure. In the former case, but *not* in the latter the values of B and S will be correlated. It is what happens in the latter case that is diagnostic for whether B causes S. The difference between observation and intervention thus roughly corresponds to the difference between so-called *back-tracking* and *non-backtracking* counterfactuals in the philosophical literature. The mark of a back-tracking counterfactual is that it involves reasoning or tracking back from an outcome to causally prior events and then perhaps forward again, as when one reasons that if the barometer reading were low (high) this would mean that the atmospheric pressure would be low (high) which in turn would mean that the storm would (would not) occur. Evaluated in this back-tracking way, the counterfactual "If the barometer reading were low (high), then the storm would (would not) occur" is true. By contrast, when the antecedent of a counterfactual is understood as made true by intervention, back-tracking is excluded, since, as emphasized above, an intervention "breaks" any previous existing relationship between the variable intervened on and its causes. Thus, when the barometer reading is set to some value by means of an intervention, one cannot infer back from this value to the value that the atmospheric pressure must have had. For this reason, the counterfactual "If the barometer reading were low (high), then the storm would (would not) occur" is *fals*e when its antecedent is regarded as made true by an intervention. Lewis holds that non-backtracking rather than backtracking counterfactuals are appropriate for understanding causation and the interventionist theory yields a similar conclusion.. This illustrates how, as claimed above, interventions play roughly the same role as the similarity metric in Lewis' theory and how they lead, as in Lewis' theory, to non-backtracking counterfactuals. with arrow-breaking having some of the features of Lewisian miracles.

What is the connection between this characterization of interventions and manipulations that are performed by human beings? I will explore this issue below but several comments will be helpful at this point. Note first that the characterization makes no explicit reference to human beings or their activities – instead the characterization is given entirely in non-anthropocentric causal language. A naturally occurring process (a "natural experiment") that does not involve human action at any point may thus qualify as an intervention if it has the right causal characteristics. Conversely, a manipulation carried out by a human being will fail to qualify as an intervention if it lacks the right causal characteristics, as in the example in which the common cause A of B and S is manipulated. Nonetheless, I think that it is plausible (section 7) that as a matter of contingent, empirical fact, many voluntary human actions as well as many behaviors carried out by animals do satisfy the conditions for an intervention. Moreover, I also think that it is a plausible empirical conjecture that humans and some other animals have a default tendency to treat their voluntary actions as though they satisfy the conditions for an intervention and to behave, learn, and in the case of humans to make casual judgments as if their learning, behavior, and judgments are guided by principles like **TC**. The connection between interventions and human (and animal) manipulation is thus quite important

to the empirical psychology of causal judgment and learning, even though the notion of an intervention is not *defined* by reference to human action.

Second, note that both **SC** and **NC** involve counterfactual claims about what would happen if certain "possible" interventions "were" to be performed. I take it to be uncontroversial that the human concept of causation is one according to which causal relationships may hold in circumstances in which it may never be within the power of human beings to actually carry out the interventions referred to in **SC** and **NC**. (In this respect the human concept may be very different from whatever underlies non-human causal cognition – section 8) Both conditions should be understood in a way that accommodates these points: what matters to whether the relationship between X and Y is causal is not whether an intervention is actually performed on X but rather what would happen to Y *if* (perhaps contrary to actual fact) such interventions *were* to be performed.

SC and **NC** connect the content of casual claims to certain counterfactuals and, as such, are not claims about how causal relationships are learned. However, if **SC** and **NC** are correct, it would be natural to expect that human beings often successfully learn causal relationships by performing interventions and in fact this is what we find. But this is *not* to say (and **SC** and **NC** do not claim) that this is the *only* way in which we can learn about causal relationships. Obviously there are many other ways in which humans may learn about causal relationships – these include passive observation of statistical relationships, instruction, and the combination of these with background knowledge. What **SC** and **NC** imply is that if, for example, one concludes on the basis of purely observational evidence that smoking causes lung cancer, this commits one to certain claims about what would happen if certain experimental manipulations of smoking were to be performed.

Finally, a brief remark about an issue that will probably be of much more concern to philosophers than to psychologists: the worry that **TC** is "circular" . Since the notion of intervention is characterized in causal terms, it follows immediately that **TC** does not provide a reductive definition of causation in terms of concepts that are non-causal. I have argued elsewhere (Woodward, 2003) that it does not follow from this observation that **TC** is uninformative or viciously circular. Rather than repeating those arguments here, let me just observe that **TC** is inconsistent with many other claims made about causation, for example, claims that causal relationships require a spatio-temporally connecting causal process. So regardless of what one makes of the "circularity" of **TC** it is certainly not vacuous or empty.

Let me now turn to the notion of direct causation. Consider a causal structure in which taking birth control pills (B) causally affects the incidence of thrombosis (T) via two different routes. B directly boosts the probability of thrombosis and indirectly lowers it by lowering the probability of an intermediate variable pregnancy (P) which is a positive cause of T (cf. Hesslow, 1976)

EMBED Word.Picture.8

Suppose further that the direct causal influence of B on T is exactly cancelled by the indirect influence of B on T that is mediated through P so that there is no overall effect of B on T . In this case B is not a total cause of T , since there are no interventions on B alone that will change T. Nonetheless, it seems clear that there is a sense in which B is a cause, indeed a direct cause, of T.

The notion of direct causation can be captured with in an interventionst framework as follows:

> (**DC**) A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set $\mathbf{V}$ is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by other independent interventions.

In the example under discussion, B counts as a direct cause of T because if we intervene to fix the value of P and then, independently of this, intervene to change the value of B the value of T will change. The notion of X's being a direct cause of Y is thus characterized in terms of the response of Y to a *combination* of interventions, including both interventions on X and interventions on other variables Z. This contrasts with the notion of a total cause which is characterized just in terms of the response of the effect variable to a single intervention on the cause variable. The notion of direct causation turns out to be normatively important because it is required to capture ideas about distinctness of causal mechanisms and to formulate a plausible relationship between causation and probabilities (for details, see Woodward, 2003, ch. 2). Of course, it is a separate question whether the notion corresponds to anything that is psychologically real in people's causal judgments and inferences. I will suggest below that it does – that it is involved in or connected to our ability to separate out means and ends in causal reasoning. It is also centrally involved in the whole idea of an intervention, which turns on there being a contrast between doing something that affects Y directly and doing something that affects Y only indirectly, through X. We will see below that even young children see able to reason causally about the consequences of combinations of interventions.

Finally let me note that both **TC** and **DC** address a very specific question: whether the relationship between X and Y is causal at all, rather than merely correlational. However if we are interested in manipulation and control, we typically want to know much more than this: we want to know *which* interventions on X will change Y, and *how* they will change Y, and under what background circumstances – that is, we want to know a whole family of more specific and fine-grained interventionist counterfactuals connecting X to Y. We may view this more detailed information, which may be captured by such devices as specific functional relationships linking X and Y , as the natural way of spelling out the detailed content of causal claims within an interventionist framework. Such information about detailed manipulability or dependency relationships is often required for tasks involving fine grained control such as tool use.

3. Additional Features of Interventionism. I said above that interventionist accounts are just one kind of approach in the more general family of theories that conceive of causes as difference makers. To further bring out what is distinctive about interventionism, consider the following causal structures.

X Y Z X Y Z

3.1 3.2

Let us make the standard Bayes' net assumption connecting causation and probabilities – the Causal Markov condition **CM** , according to which conditional on its direct causes, every variable is independent of every other variable, singly or in combination, except for its effects. Given this assumption, both structures 3.1 and 3.2 imply exactly the same conditional and unconditional independence relationships: in both X, Y and Z are dependent and X and Z are independent conditional on Y. The difference between the structures 3.1 and 3.2 shows up when we interpret the directed edges in them as carrying implications about what would happen if various hypothetical interventions were to be performed, in accordance with **DC**. In particular, if 3.1 is the correct structure, under some possible intervention on Y, X and Z will change, while if 3.2 is the correct structure Z but not X will change under an intervention on Y. Similarly, 3.2 implies that under some intervention on X, both Y and Z will change, while 3.1 implies that neither Y nor Z will change. In general, if two causal structures differ at all, they will make different predictions about what will happen under some hypothetical interventions, although, as 3.1-3.2 illustrate, they may agree fully about the actual patterns of correlations that will be observed, in the absence of these interventions.

Although an interventionist account does not attempt to reduce causal claims to information about conditional probabilities, it readily agrees that such information can be highly relevant as evidence for discriminating between competing causal structures. Indeed, as explained in Woodward, 2003, pp. 339ff, we may think of **CM** as a condition that connects claims about what happens under interventions to claims about conditional probabilities involving observed outcomes, thus allowing us to move back and forth between the two kinds of claims. Arguably (see Section 8) the ability to move smoothly from claims about causal structure that follow from information about the results of interventions to claims about causal structure that are supported by observations and vice-versa is one of the distinctive features of human causal cognition. In this connection, there is considerable evidence that at least in simple cases humans can learn causal Bayes nets from passive observations, interventions and combinations of the two. Indeed, for at least some tasks the assumption that subjects are Bayes' net learners does a better job of accounting for performance than alternative learning theories.

I suggested above that an interventionist account will lead to different causal judgments about particular cases than causal process accounts. Consider cases of "double prevention" in which A prevents the occurrence of B which, had it oc-

curred, would have prevented the occurrence of a third event C, with the result that C occurs. Cases of this sort occur in ordinary life and are common in biological contexts. For example, the presence (A) of lactose in the environment of E. Coli results in the production (C) of a protein that initiates transcription of the enzyme that digests lactose by interfering with the operation (B) of an agent that (in the absence of lactose) prevents transcription. There is dependence of the sort associated with interventionist counterfactuals between whether or not lactose is present and the synthesis (or not) of the enzyme which digests it – manipulating whether lactose is present changes whether the enzyme is synthesized -- but no spatio-temporally continuous process or transfer of energy, momentum, or force between lactose and the enzyme. Interventionist accounts along the lines of **TC** will judge such relationships as causal while causal process theories will not. Biological practice seems to follow the interventionist assessment, but it would be useful to have a more systematic experimental investigation of whether ordinary subject regard double prevention relationships as causal, how they assess causal efficacy or strength in such cases, and the ease with which such relationships can be learned.

Double prevention cases suggest that energy transmission is not necessary for causal relatedness. Is it sufficient? Arguably, energy transmission between two events is sufficient for there being *some* causal process connecting the two. However, the information that such a process is present is *not* tantamount to the detailed information about dependency relationships provided by interventionist counterfactuals. This is suggested by the following example. (Hitchcock, 1995). A cue stick strikes a cue ball which in turn strikes the eight ball causing it to drop into a pocket. The stick has been coated with blue chalk dust, some of which is transmitted to the cue ball and then to the eight ball as a result of the collision. In this case, energy, momentum, and "force" are all transmitted from the stick to the cue ball. These quantities are also transmitted through the patches of blue chalk that eventually end up on the eight ball. The sequence leading from the impact of the cue stick to the dropping of the eight ball is a causal process, as is the transmission of the blue chalk, and a connecting mechanism is present throughout this sequence. The problem is that there is nothing in all this information that singles out the details of the way in which cue stick strikes the cue ball (and the linear and angular momentum that are so communicated) rather than, say, the sheer fact that the cue stick has struck the cue ball in some way or other or the fact that there has been transmission of blue chalk dust as causally relevant to whether the eight ball drops. Someone might fully understand both the abstract notion of a causal process and be able to recognize that the process connecting cue stick, cue ball and eight ball is a causal process that transmits energy and yet not understand how variations in the way the cue strikes the cue ball make a difference to the subsequent motion of the eight ball, and that the transmission of the chalk dust is irrelevant. Yet this information, which is captured by interventionist counterfactuals of the sort described in **TC**, is crucial for manipulating whether the eight ball drops in the pocket. As we will see, this observation has implications for primate causal

understanding.

In general, then, an interventionist account predicts that when information about spatio-temporal connectedness is pitted against information about dependency relations of the sort captured by interventionist counterfactuals, the latter rather than the former will guide causal judgment. For example, if the relationship between C and E satisfies the conditions in **TC**, people will judge that C causes E even if there appears to be spatio-temporal gap between C and E. Moreover, even if there is a connecting spatio-temporally continuous process from C to E, they will judge that C does not cause E if the dependence conditions in **TC** are not satisfied. Similarly, for the information that something has been transmitted from C to E: although chalk dust is transmitted to the eight ball, subjects will not judge that its presence causes the ball to go into the pocket because the conditions **TC** are not satisfied.

Despite these observations, adherents of an interventionist account can readily acknowledge that information about causal mechanisms, properly understood, plays an important role in human causal learning and understanding. However, rather than trying to explicate the notion of a causal mechanism in terms of notions like force, energy, or generative transmission, interventionists will instead appeal to interventionist counterfactuals. Simplifying greatly, information about a mechanism connecting C to E will typically be information about a set of dependency relationships, specified by interventionist counterfactuals, connecting C and E to intermediate variables and the intermediate variables to one another, perhaps structured in a characteristic spatio-temporal pattern (cf, Woodward, 2002). Among other things, such counterfactuals will specify how interventions on intermediate variables will modify or interfere with the overall pattern of dependence between C and E. As an illustration, consider Shultz's classic, 1982 monograph in which he argues that children rely heavily on mechanism information in causal attribution. This mechanism information can be readily reinterpreted as information about interventionist counterfactuals. For example, in experiment two, subjects must decide which of two different lamps is responsible for the light projected on a wall. Here the relevant interventionist counterfactuals will describe the relationship between turning on the lamp and the appearance of a spot on the wall, the orientation of the lamp and the position of the spot, the effect of inserting a mirror in the path of transmission, and so on. Similarly, in the cue ball example, the relevant mechanism will be specified in terms of the dependence of the trajectories of the cue and eight ball on variations in the momentum communicated by the stick, the effect of intervening independently on the eight ball (e.g. gluing it to the table) and so on.

On this construal, detailed information about the operation of mechanisms is not, as is often supposed, something different in kind from information about dependency or manipulability relationships, understood in terms of interventionist counterfactuals, but rather simply more of the same: more detailed fine grained information about dependency relationships involving intermediate variables.

An additional advantage of this way of looking at things is that it provides a natural account of how it is possible, as it clearly is, for people to learn that there is a causal relationship between C and E without knowing anything about a connecting mechanism. This is much harder to understand if, as some mechanism –based approaches claim, the existence of a causal relationship between C and E just consists in the obtaining of a connecting mechanism between C and E and the information that C causes E consists in or implies information to the effect that there is such a mechanism. By contrast, according to **TC**, people will judge that C causes E, if they are presented with evidence (e.g. from repeated experimental manipulations) that the relevant interventionist counterfactuals hold between C and E, even if they have no information about an intervening mechanism.

4. Philosophy and psychology. The interconnections between philosophical and psychological treatments of causation are complex and intricate. Many although by no means all philosophical accounts are (at least officially) intended as accounts about the world rather than as accounts of anyone's psychology: that is, as accounts of what causation *is* or (less ambitiously) of constraints that hold between causal relationships, as they exist in the world, and other wordly relationships (having to do, e.g. with the obtaining of regularities). Nonetheless, it is common for philosophers to move back and forth between such wordly claims and claims that do sound more psychological: claims about what people mean (or ought to mean) when they make causal claims, the evidence on which such claims are or should be based and so on. Moreover, even when no such accompanying psychological story is explicitly described, it is often implicit in or at least naturally suggested by the ostensibly wordly account. For example, it is natural to suppose that philosophers who claim that causation can be reduced to facts about conditional probabilities will also think that human causal beliefs and representations encode facts about conditional probabilities, and that causal learning consists in learning facts about conditional probabilities. Similarly, if a theorist claims, as some adherents of causal process/mechanistic approaches do, that whether C causes E has nothing to do with what does or would happen to E in the absence of C, one would not expect (at least on the face of things) human causal judgment to represent or to be sensitive to such information.

Matters are further complicated, though, by the fact that insofar as philosophical accounts of causation have psychological implications, they are often presented primarily as *normative* rather than straightforwardly *descriptive* accounts – that is, they are presented as accounts of the causal judgments people ought to make in various situations, how they ought to use evidence in reaching such judgments, and so on. I assume, however, that it is always in order to ask how these accounts fare when taken as descriptive theories – we may construe them as descriptive claims, regardless of the intentions of their authors. Moreover, quite apart from its great intrinsic interest, there is an obvious motivation for proceeding in this way. Humans and other animals engage in a remarkable amount of *successful* causal learning and form many true or correct causal representations of the world. There must be some unified story about this that is

both an accurate description of what they do and that enables us to understand how what they do leads, often enough, to normatively correct outcomes. Asking about the descriptive adequacy of various normative theories is an obvious route to this sort of understanding.

In addition, there are many other interconnections between normative and descriptive theories. It is very common for philosophers to appeal both to claims about the causal judgments that ordinary people or experts will make in particular cases and to claims about the kinds of considerations on which those judgments are based to motivate the particular theories they favor. It is also common for philosophers to make claims about how people's causal judgments connect up with or fail to connect up with various other concepts and patterns of reasoning such as the use of counterfactuals in order to motivate particular approaches. Claims of this sort are of course descriptive claims about the empirical psychology of causal inference and judgment and should be evaluated accordingly. In addition, while adherents of a normative theory always have the option, in any particular case, of responding to evidence that subjects do not in fact reason and judge in the way that theory says they should, by saying that such subjects are subject to processing limitations, or are confused, extensive and fundamental divergence between normative prescriptions and actual behavior is often plausibly regarded as at least a prima-facie problem for a normative theory – a problem that the normative theory needs to address rather than ignore. In the spirit of these remarks, I will explore, in the remainder of this paper, some issues concerning the empirical plausibility of interventionist accounts and their philosophical rivals as descriptions of human and non-human causal inference and judgment.

5. Instrumental Learning. A useful point of departure is the difference between, on the one hand, classical or Pavlovian conditioning and, on the other, instrumental or operant conditioning. In classical conditioning, a subject learns an association between two events that are outside of its control – e.g. an association between the ringing of a bell and the provision of food. The subject is thus in the position of learning through passive observation rather than active intervention and what is learned is that one stimulus predicts another, where this predictive relationship may or may not reflect the fact that the first stimulus causes the second. By way of contrast, in instrumental conditioning what is learned is an association between some behavior produced by the subject and an outcome, as when rats learn an association between pressing a lever and the provision of a food pellet.

From an intervenionist perspective, instrumental learning has a "cause-like" flavor. An organism that was incapable of acting on the world and could only passively observe associations outside of its control would have no need for a notion of causation or cause-like representations, conceived along interventionist lines. Such an organism might still find it useful to predict what will happen but sensitivity to correlations and to temporal relationships, rather than to anything distinctively causal would suffice for this purpose. Given a correlation

between two variables, X and Y, it would not matter how the correlation arises – whether because (i) X causes Y or because (ii) X and Y have a common cause --as long as the correlation is stable and projectable. The difference between (i) and (ii) begins to matter when the animal is interested in whether changing X is a way of changing Y.

It is thus of considerable interest that there are striking, if incomplete, parallels between instrumental conditioning in non-human animals and causal learning and judgment in humans – a theme that has been systematically explored by Dickinson, Shanks and others in a series of papers (Dickinson and Shanks, 1995, Dickinson and Balleine, 2000). Both instrumental learning by rats and human judgments of causal strength (as expressed in verbal reports) in instrumental learning tasks exhibit a similar sensitivity to temporal delay between action and outcome. Both rat behavior and human causal judgment are (independently of temporal relations) highly sensitive to the *contingency* Δp between action A and outcome O – that is, to *P(O/A)- P(O/-A)*. Although there are important qualifications, both human judgments of the causal strength and the rate of lever pressing for rats tend to decline as Δp approaches zero. In addition, in both humans and rats, learning of instrumental contingencies has a number of other features that gives it a causal flavor –for example, both exhibit backward blocking and both rat behavior and human causal judgment are subject to a discounting or signaling effect in which the usual reaction of non-response to a non-contingent reward schedule does not occur when rewards that are not paired with the instrumental action are preceded by a brief visual signal, As Dickinson and Balleine remark, "the intuitive explanation [of this effect] is that the signal marks the presence of a potential cause of the unpaired outcomes, thereby discounting these outcomes in the evaluation of control exerted by the instrumental action" (2000, p. 192).

These results suggest that both instrumental learning in rats and human judgments of causal strength (as well as actions based on this) behave as though they track the perceived degree of control or manipulative efficacy of the instrumental action over the outcome, which is what one would expect on an interventionist account on causation. In addition, phenomena such as sensitivity to contingency, backward blocking, and causal discounting show that at least some causal representation and judgment is sensitive not just to information about the rates of occurrence of cause and effect and the processes that connect them but also to information about what would or does happen in the absence of the cause, and under the occurrence of potential alternative causes of the effect. This is contrary to what some (psychologized) versions of causal process/mechanism theories seem to imply.

6. Causal Judgment and Interventionist Counterfactuals. I noted above that interventionist theories are just one species of the more general category of difference-making theories. The sensitivity of causal judgment to contingency information is consistent both with various versions of probabilistic theories of causation as well as with theories that appeal to interventionist counterfactuals.

Is there evidence that specifically favors interventionism as a descriptive account of causal judgment, at least in humans?

Let me begin with the issue of the relationship between causal and counterfactual judgments. Although, as noted above there are influential philosophical theories such as Lewis, 1973 that connect causal claims to counterfactuals many philosophers continue to regard counterfactuals in general (and *afortiori*, their use in a theory of causation) with great skepticism. It is contended that counterfactuals are unclear, untestable, unscientific, and in various ways unnatural and artificial in the sense that they are philosophical inventions that correspond to nothing in the way ordinary people actually think and reason.

In fact there is considerable evidence that people employ counterfactuals extensively in various forms of ordinary reasoning and that they connect causal claims and counterfactuals in something like that interventionist and counterfactual theories suggest. Since the relevant literature is vast, I will focus, for illustrative purposes, on a charming set of experiments involving young children described in Harris (2000). Harris presented children aged 3-4 with a number of scenarios which probed the way in which they connected causal and counterfactual judgments. He found, for example, that when children were presented with a causal sequence (Carol walks across the floor in her muddy shoes and makes the floor dirty) and then asked counterfactual questions about what would have happened under different possible antecedents (what would have happened if Carol had taken her shoes off), a large majority give correct answers (that is, answers that respect the intuitive connection between causal and counterfactual claims). They are also able to discriminate correctly between counterfactual alterations in the scenario that would have led to the same and to different outcomes- i.e. which alterations in behavior would have avoided mud on the floor and which would not.

Children do not connect causal and counterfactual claims only when explicitly prompted to do so by a question about what would happen under a counterfactual possibility; they also do this when asked why an outcome occurred or how it might have been prevented. For example, in a scenario in which Sally has a choice between drawing with a pen and drawing with a pencil, chooses the pen, and gets ink on her fingers, children who are asked why Sally's fingers got inky motivate the causal role of the pen by appealing to what would have happened if she had instead used the pencil. Indeed, children spontaneously invoke what would have happened under alternative possibilities in arriving at causal judgments even when those alternatives are not explicitly mentioned in or prompted by the scenarios they are given. Harris' conclusion is that "counterfactual thinking comes readily to very young children and is deployed in their causal analysis of an outcome" (136).

This conclusion may seem surprising if one is accustomed, as many philosophers are, to thinking of counterfactuals as primarily having to do with Lewis –style similarity relationships on possible worlds and similar metaphysical arcana. Clearly small children (and for that matter most adults) don't have anything

remotely like Lewis' framework explicitly in mind when they use counterfactual reasoning. But whatever one's assessment of Lewis' theory, it is important to bear in mind that one of the main everyday uses of counterfactual and causal thinking, by both children and adults, is in planning and in anticipating what the consequences of various possible courses of action would be (without necessarily performing the actions in question). This is a perfectly ordinary, natural, practically useful activity and (relevantly to our story – see below) one that even small children appear to be much better at than non-human primates. Children engage in such planning involving counterfactuals and causal claims on an everyday basis when they reason, for example, that if they want to avoid getting their fingers inky they should use a pencil rather than a pen, that using a pen with blue ink rather than black ink will not avoid the outcome and so on. If we think of counterfactuals of this sort, used for this purpose (notice, by the way, that the above counterfactuals are all interventionist counterfactuals) we should be able to see that there is nothing particularly problematic or obscure about them.

Turning now specifically to the notion of an intervention, a natural worry is that this notion is too complex and cognitively sophisticated to be psychologically realistic. In assessing this worry we need to distinguish two issues: 1) Do most people consciously or explicitly represent to themselves the full technical definition of a normatively appropriate notion of intervention when they engage in causal reasoning? 2) Do people learn and reason in accord with the normative requirements of the interventionist account? I assume that the answer to 1) is almost certainly, "no", for most people without special training. On the other hand, there is considerable evidence that the answer to 2) is, "yes, for many people at least some of the time".

To begin with, there is evidence that in a substantial range of situations adults learn causal relationships more reliably and quickly when they are able to perform interventions than when they must rely entirely on passive observations (Lagnado and Sloman, 2004, Sobel and Kushnir, 2003) . This true for infants as well—Jessica Sommerville (this volume) reports a series of experiments that show that infants who actively intervene to e.g., obtain a toy by pulling a cloth on which it rests learn to distinguish relevant causal relationships between the cloth and toy (presence of spatio-temporal contact etc.) more readily than those who rely on passive looking. Moreover, in at least some situations a significant number of subjects (although by no means all) intervene optimally when given a choice among which interventions to perform, choosing those interventions that are maximally informative. For example, when presented with a scenario in which there are several possible candidates for the correct causal structure, one of which is a chain structure in which X causes Y which causes Z, people choose to intervene on the more diagnostic intermediate variable Y, rather than on X or Z . (Steyvers, Tenenbaum, Wagenmakers, and Blum, 2003) This suggests some appreciation of the connection between intervention and causal structure.

A similar conclusion is suggested by a series of experiments by Lagnado and

Sloman (Forthcoming, Sloman and Lagnado, 2004) They report the following:

6.1) Subjects are told that billiard ball one causes ball two to move which causes ball three to move. Almost all judge that if ball two were unable to move, ball one would still have moved, and that billiard ball three would not have. On other hand, when presented with a parallel scenario involving conditionals that lack an obvious causal interpretation and are of the form if p then q, if q then r, subjects' responses are far more variable, with a considerable number willing to infer not p from the information that not q. In another words, most subjects endorse the non-backtracking counterfactuals associated with interventionist accounts in the causal scenario but respond differently to non-causal conditionals, where a considerable number do endorse a backtracking, non-interventionist interpretation.

6.2) Subjects are presented with a chain structure in which they are told that A causes B which causes C. They are then told either (a) "someone intervened directly on B, preventing it from happening" or (b) we "observe" that B didn't happen. Again consistently with the interventionist account, subjects treat the "intervention" condition (a) very differently from the "observation" condition (b). For example, they judge that probability of A is higher in the intervention condition than in the observation condition -- that is they don't backtrack in the former, and are more likely to in the latter.

These and other experiments involving more complex causal structures suggest that subjects do indeed distinguish between observing and intervening in the way that the interventionist account says they should, that in at least some situations they interpret an intervention in an arrow-breaking way, and that they associate interventionist non-backtracking counterfactuals with causal claims and employ them in contexts in which a causal interpretation is natural or a reasonable default, while being at least somewhat more inclined to use non-backtracking counterfactuals in contexts that are obviously non causal. These results seem inconsistent with claims (e.g. Bennett, 1984) in the philosophical literature that people either do not distinguish at all between backtracking and non-backtracking counterfactuals, or do not preferentially employ the latter in contexts involving causal reasoning . In addition, the experiments provide additional evidence (if any is needed) that subjects are indeed able to engage in sophisticated normatively appropriate counterfactual reasoning regarding causal situations.

7. Interventions and Voluntary Actions. I noted above that in many situations people make more reliable causal inferences when they are able to intervene. From a design viewpoint, one thus might expect that subjects will have more confidence in causal inferences and judgments that are directly associated with their interventions and perhaps that some of these inferences will be fairly automatic. This suggests the following hypothesis: human beings (and perhaps some animals) have (i) a default tendency to behave or reason as though they take their own voluntary actions to have the characteristics of interventions and (ii) associated with this a strong tendency to take changes that temporally

follow those interventions (presumably with a relatively short delay) as caused by them. "Voluntary" here means nothing metaphysically fancy– just the common sense distinction between deliberately pouring the milk in one's coffee and spilling it accidentally.

I noted above that it is not psychologically realistic to suppose that most people operate with an explicit representation of the full technical definition of the notion of an intervention. Taken together (i) and (ii) suggest one way in which it is nonetheless possible for such subjects to use their interventions (note: not their explicit concepts of intervention) to fairly reach reliable causal conclusions in way that respects principles like (**TC**). For an account along these lines to work several things must be true. First, subjects must have some way of determining (some signal that tells them) when they have performed a voluntary action and this signal must be somewhat reliable, at least in ordinary circumstances. Second, voluntary actions (again in ordinary, ecologically realistic circumstances) must -- not always, but often enough -- have the characteristics of an intervention. I suggest that both claims are true. First, human subjects do have a characteristic phenomenology which is associated with voluntary action -- they typically have a sense of agency or ownership of their behavior that is not present when they act involuntarily. This is not surprising: presumably it is very important for humans and other animals to have some way of distinguishing those cases in which a change occurs in their environments or in their bodies that results from their voluntary actions from those cases in which the change comes about in some other way – not as a result of a movement of their bodies at all, or as a result of a movement that is non-voluntary. It is plausible that one role for the feeling of ownership of one's action is to provide information that helps organisms to monitor this distinction. Once this feeling is available, it may be used for many purposes, including causal inference.

Turning now to the status of (ii), it is clear that the correlation between voluntariness and satisfaction of the conditions for an intervention is imperfect. In a badly designed clinical trial, an experimenter might be subconsciously influenced, in his decisions to give a drug to some patients and withhold it from others, by the health of the patients—his decisions are "voluntary" and yet correlated with an independent cause of recovery in a way that means that the conditions for an intervention are not satisfied. Nonetheless, it seems plausible that many voluntary actions do, as a matter of empirical fact, satisfy the conditions for an intervention. If I come upon a wall switch in an unfamiliar house and find that there is a regular association between my flipping the position of the switch and whether a certain overhead light is on or off, then often enough my flippings will satisfy the conditions for an intervention on the position of the switch with respect to the state of the light. Similarly for a baby whose leg is attached by a string to a mobile and who observes a correlation between her leg movements and the motion of the mobile. In both of the cases, subjects who are guided by (i) and (ii) above will make fairly reliable causal inferences. The existence of causal illusions in which we experience or "perceive" salient changes that follow our voluntary actions as caused by them similarly suggests that such

a heuristic is at work. Going further, it might be conjectured that involuntary behavior is less likely to meet the conditions for an intervention. If this is so, one might expect that the impression. of causal efficacy for outcomes following such behavior should be attenuated. Premack and Premack, 2003 report this is the case, although more systematic experimental investigation would be desirable.

8. **Primate Causal Cognition**. Despite the abilities of non human animals in instrumental learning tasks and the similarities between animal instrumental and human causal learning described above, it is a striking fact that non-human animals, including primates, are greatly inferior to humans, including small children, at many tasks involving causal learning, especially those involving tool use, object manipulation, and an understanding of "folk physics". This is so despite the fact that non-human primates and many other mammals have capacities on object permanence and trajectory completion tasks (capacities that are often taken to demonstrate the possession of "causal" concepts in the psychological literature) that that are apparently not so very different from those possessed by human children and adults. This suggests that while these various abilities may well be necessary for the acquisition of the causal learning abilities and understanding possessed by human beings they are not sufficient. Can an interventionist perspective cast light on what more is involved?

In approaching this question, let me begin by briefly describing some representative experimental results involving non-human primates. In experiments conducted by Kohler and subsequently repeated by others, apes (including chimps, orangutans and gorillas) were presented with problems that required stacking several boxes on top of each other in order to reach a food reward. In comparison with humans, including children, the apes had great difficulty. They behaved as though they had no understanding of the physical principles underlying the balancing of the boxes and the achievement of structures capable of providing stable support – as Kohler put it, they had "practically no statics" (Kohler, 1927, p. 149, quoted in Povinelli, 2000, p. 79). The structures they succeeded in building, after considerable trial and error, were highly unstable, and completely neglected center of gravity considerations, with boxes at an upper level extending in a haphazard way far over the edges of lower level boxes. Subjects even on occasion removed lower level boxes from beneath boxes they supported. Errors of this sort were made repeatedly, suggesting what from a human perspective would be described as complete lack of insight into the principles governing the construction of stable structures. When stable structures were achieved, this appeared to be the result of trial and error learning. There was little evidence that the apes were able to reason hypothetically about what would happen if they were to create this or that structure, without actually creating the structures in question, and then use this reasoning to guide their actions in the way that, e.g., the children in Harris' experiments were able to reason.

In another series of experiments, conducted by Visalberghi and Trinca (1989), a desirable food item was placed in a transparent hollow tube and the animals were

given various tools that might be used to push it out. Both apes and monkeys were able to solve some variants of this problem. For example, when given a bundle of sticks that was too thick to fit into the tube, they unbundled the sticks and used appropriately sized sticks to dislodge the food item. On the other hand, they also frequently behaved as though the lacked a real understanding of the causal structure of the task. For example, they inserted sticks that were too short to reach the reward when a stick of appropriate length was available. They attempted to use sticks with cross pieces that blocked insertion into the tube. They also inserted non-rigid objects like tape that were incapable of displacing the food. In still other experiments, the animals failed to choose implements with a hook at the end, which would have been effective in retrieving desired objects instead of straight sticks, which were not.

Povinelli's summary is that the animals "appear to understand very little about why their successful actions are effective". (2000, p 104). In particular, they appeared to not to understand the significance of the mechanical properties of the systems they were dealing with – properties such as weight, rigidity, shape, center of mass, and so on. Instead, as both Povinelli and Tomasello and Call remark, they often acted as though (any) spatio-temporal contact between the target object they wished to manipulate and the means employed was sufficient to achieve the desired manipulation.

Both Povinelli, 2000 and Tomasello and Call, 1997 go on to suggest a more general characterization of the deficits exhibited in the experiments: they claim that these stem from the animals' lack of various abstract concepts having to do with "unobservables" (Povinelli, 2000, p. 300 mentions gravity, force, shape and mass, among others) that humans think of as mediating causal relationships. In contrast to humans, apes operate entirely within a framework of properties that can be readily perceived and this underlies their lack of causal understanding.

Philosophers of science are likely to find this invocation of "unobservables" puzzling. If we think of a property as "observable" for a subject as long as the subject can reliably discriminate whether or not it is present (or among different values if the property is quantitative) by perceptual means, then it seems implausible that properties like weight and shape are literally unobservable by apes- presumably, apes can be trained to reliably discriminate between objects of different shapes or weights. There is, however, an alternative way of understanding this claim that makes it seem far more plausible. Suppose that when an ape learns to discriminate among objects according to (what we would call) weight the discrimination is made on the basis of sensory feedback and bodily sensations associated with differential effort in lifting. If apes' "concept" of weight is very closely linked to these bodily sensations, then it becomes more understandable why they are apparently unable to make use of information about weight in other sorts of contexts requiring causal reasoning- why, for example, they are unable to recognize the relevance of weight to support relationships. To recognize the relevance of weight to these contexts requires possession of a more abstract way of thinking about weight that is not so closely tied to sen-

sory and motor experience. Similarly for properties like rigidity. On this way of thinking about the matter, the apes e (in comparison with humans) operate with the wrong variables to enable them to engage in the kind of sophisticated causal learning required for the tasks described above- their variables are too closely linked to egocentric sensory experience. From the perspective of the interventionist account, we might describe this as a situation in which certain interventionist counterfactuals cannot be learned by the apes because the variables in terms of which those counterfactuals are framed are unavailable to the apes. For example, apes are unable to learn the appropriate interventionist counterfactuals involving the human concept of weight because they lack that concept. Whether or not this analysis is accepted, it seems clear, as a more general point, that whatever the apes' grasp of notions like weight and rigidity, they do not understand their causal relevance to the tasks with which they are dealing and cannot integrate these notions into causal representations that successfully guide action in connection with those tasks.

As I see it, this sort of limitation on the apes' understanding is not just a matter of their failure to grasp the abstract notion of a causal process (as a process that transmits force, energy etc.) or an inability to recognize particular instances of such a process in thes system of interest. As noted in Section 3, grasp of the notion of a causal process is *not* sufficient for the sort of detailed knowledge of dependency relationships that is required for successful manipulation in tasks like balancing boxes or extracting food from a tube. What needs to be explained is the apes' lack of this latter sort of knowledge. Whenever a primate moves a food source with a stick – whether it is pushed in an appropriate or inappropriate direction or with an appropriate instrument-- there will be transmission of "force" and energy, the presence of a mechanism and so on. A creature which possessed the concept of force and "generative transmission" (and which could recognize when force was being transmitted) and whose heuristic was: to cause a desired outcome transmit force to the outcome (or the object associated with the outcome) or set in operation a generative mechanism connected to the outcome would not get useful guidance from this heuristic about exactly what it should do to balance boxes in the stacking task or to expel food from the tube. To accomplish this, far more specific information about how the outcome that the agent wishes to affect depends on variation in other factors (perhaps including factors that are not linked too closely to egocentric sensory experience) that the agent is able to control is required, where these includie factors that are not linked too closely to egocentric sensory experience: thus in the tube experiment the subject must recognize the relevance of the dimensions and rigidity of the implement chosen and so on. This looks far more like information of the sort represented by **TC** and **DC** (see below) than information about force transmission.

The idea that the apes lack the right variables (and hence cannot grasp counterfactual dependency relationships based on those variables) gives us one way of explaining at least some of their deficits in causal understanding. An alternative line of argument, which I see as complimentary to and not in competition with

the "wrong variables" analysis, and which also fits naturally into an interventionist framework focuses on Tomasello's and Call's notion of a *tertiary* relationship. (1997, especially pp 367- 400). A relationship qualifies as tertiary for a subject if the relationship is understood or recognized as holding between objects and individuals that are independent of the subject. This contrasts with relationships that are (or are conceived of as) more directly egocentric in the sense of holding between the subject and some other object or individual. Clearly, the ability to recognize and reason in terms of tertiary relations is closely related to the ability to think in an abstract or context-independent way. Tomasello and Call suggest that all primates (or ar least all simians) have the ability to form and understand concepts of tertiary relationships in both social and physical domains. For example, primates seem to possess concepts of tertiary social relationships between conspecifics, such as the concept of one animal out-ranking another in a dominance hierarchy (as opposed to the notion of the non-tertiary relationship of this animal outranking me).

This suggests the following question: do primates understand (or at least behave in accordance with a conception of) causation as a tertiary relationship? As argued in section 2, the human concept of causation is clearly a concept of a tertiary relationship. Although people think of causal relationships as relationships that they may be able to exploit for purposes of manipulation and control, they also conceive of causal relationships as relationships that can exist in nature independently of their (or indeed any agent's) manipulative activities. Thinking along these lines suggests the usefulness of distinguishing among the following possibilities or "levels" of causal/instrumental understanding:

> 8.1) First, consider an agent creature whose instrumental behavior and learning is purely *egocentic.* That is, the agent grasps (or behaves as if it grasps) that there are regular, stable relationships between *its* manipulations and various downstream effects but stops at this point, not recognizing (or behaving as though it recognizes) that the same relationship can be present even when it does not act, but other agents act similarly or when a similar relationship occurs in nature without the involvement of any agents at all.
>
> 8.2) Second, consider an agent with an *agent causal* viewpoint: the agent grasps that the very same relationship that it exploits in intervening also can be present when other agents act .
>
> 8.3) Third, consider an agent with a *fully causal* viewpoint: the agent grasps that the very same relationship that he exploits in intervening also can be present both when other agents intervene and in nature even when no other agents are involved. This involves thinking of causation as a tertiary relationship.

Tomasello and Call suggest that non-human primates do not operate with this tertiary, stage three conception of causation but rather with something closer to what I take to be the egocentric conception described in 8.1:

> we are not convinced that apes need to be using a concept of causality in the experimental tasks purporting to illustrate its use, at least not in the humanlike sense of one independent event forcing another to occur. More convincing would be a situation in which an individual observes a contiguity of two events, infers a cause as intermediary, and then finds a novel way to manipulate that cause. For example, suppose that an individual ape, who has never before observed such an event, for the first time observes the wind blowing a tree such that the fruit falls to the ground. If it understands the causal relations involved, that the movement of the limb is what caused the fruit to fall, it should be able to devise other ways to make the limb move and so make the fruit fall. … we believe that most primatologists would be astounded to see the ape, *just on the bases of having observed the wind make fruit fall*, proceed to shake a limp, or pull an attached vine, to create the same movement of the limb. Again, the problem is that the wind is completely independent of the observing individual and so causal analysis would have to proceed without references to the organism's own behavior and the feedback it might receive from that (thus, it might be able to learn to shake the limb if its own movements had previously led to a limb shaking and the fruit falling as a result). Moreover, performing some novel behavior to make the fruit fall would involve an even deeper causal analysis of the web of possible ways that the cause could be repeated so as to reinstate the desired effect.

EMBED MSDraw.Drawing.8.2

(1997, p 389).

Although some commentators (e.g. Povinelli, 2000) are skeptical, I think that these remarks help to capture some important features of the limitations exhibited in the primate experiments described above. In what follows I want to develop some of the implications of this line of thought in more detail.

First, note that the transitions from levels 1-3 are important in part because correspond to progressively stronger forms of instrumental/causal learning. For example, if I am a creature who thinks only in terms of instrumental relationships that connect my own actions to outcomes (level 1 above) and not in terms of levels 2 and 3, then the relevance of observations concerning what happens under the interventions of others will be unclear to me. Suppose that I do X, and observe that Y ensues and that I have the ability to learn, from repeated experiences of this sort, that (usually or often) when I do X, Y regularly ensues. Clearly, it is logically possible that I might have this ability and yet not be able to learn or recognize that if when another actor does X, Y ensues, then this is evidence that if I were to do X, Y would ensue. Similarly, I may have the level 1 learning ability just described and not be able to recognize that there are relationships that occur in nature in the absence of human or animal intervention that are such that I could make use of those very relationships for purposes of

manipulation . Associated with this, I may not be able to learn from observing naturally occurring events such as that these instantiate relationships that I myself might make use of for purposes of manipulation. In short, in level 1, the only way I learn about a manipulative relationship is if I myself perform the relevant manipulation. I take Tomasello and Call to be suggesting that this is not just a logical possibility but that something like this is true for non-human primates.

This line of thought suggests that susceptibility to instrumental conditioning shows only that an animal is capable of learning instrumental relationships in the sense of level 1; it does not in itself show that the animal is capable of understanding or appreciating causal relationships in sense 3 or the forms of learning associated with it. What would go some way (see below) toward establishing the latter would be evidence of transference between operant and classical conditioning. Suppose that C is some outcome that an animal knows how to produce and that the animal learns that C is associated with E just via passive observation or classical conditioning where E is an outcome the animal wants. Will the animal spontaneously produce C (without extensive trial and error learning) in order to get E once it is given the opportunity to intervene? If the animal learns in an instrumental conditioning task that producing C is followed by E will the animal expect (or quickly learn to expect) E when it merely observes but does not produce C? Although there is some controversy surrounding this issue, the consensus seems to be that there is relatively little transfer back and forth between instrumental and classical conditioning.This is consistent with the claims of Tomasello and Call about the inability of non human primates to learn instrumental relationships from passive observation of causal relationships occurring in nature. If correct, such claims do indeed suggest that the representations and abilities that underlie non-human instrumental learning are not fully causal in the human, level three sense, even though as indicated earlier they have many features in common with human causal learning and representation.

There is another aspect of the contrast between stages 1 and 3 that is worth underscoring. An animal which possess only stage 1 information is in effect in the position of possessing fused action-outcome representations and behavior patterns: representations that its behaving a certain way produces such and such a desired outcome/goal. This need not involve any appreciation of causal relationships among varaibles that are intermediate between the behavior and achievement of the goal. It thus falls well short of what might be thought of as full-fledged means/ends understanding of how the goal might be achieved. This last does involve the postulation of intermediate causal links or what I take to be the same thing, some appreciation of the contrast between direct and more indirect casual relationships. In particular, means/ends understanding seems to involve a decomposition of a task into an intermediate outcome O that can be produced fairly directly by the subject's action A and a further outcome O' that is more directly caused by O and less directly by A, and where the link between O and O' is a tertiary link between events, rather than an action-event link. In Tomasllo's and Call's diagram, this intermediate outcome is described

by the variable "limb shakes" and this in turn causes the outcome described by "fruit falls". Note that this causal relationship holds between events neither of which is a manipulation by the animal. As Tomasello and Call suggest, and is apparent from their diagram, it is the introduction of the intermediate variable that makes possible (or corresponds to) the recognition that there are different ways (involving both actions and events occurring in nature) in which the same goal (fruit falls) might be brought about, all of which have in common the fact that they operate through the intermediate variable "limb shaking". In general, the postulation of the intermediate link (and with it an appreciation that causal relationships can be more or less direct) goes hand in hand with a decoupling of sought after final outcomes and the means used to achieve them and a focus on the latter as a separate entity.

As Tomasello and Call, 1997 and Tomasello, 1999 argue, this decoupling is closely linked to learning through imitation – that is, through observing the interventions of others. The issue of whether non-human primates ever learn through genuine imitation, as opposed to such other possibilities as emulation learning, is a complex and controversial one, involving , among other things, disputes over how to best characterize imitation and issues about the theory of mind skills required for this activity. However, it seems uncontroversial that in comparison with humans, including young children, non-human animals, including primates, are much inferior at learning means/ends relationships and appropriate tool use by observing the manipulations of other conspecifics. It also seems uncontroversial that whatever else is required for successful imitation, the ability to perform the kind of means/ ends (goal) decomposition described by Tomasello and Call is essential. One reason for thinking this is that if imitation is to be successful it will often not involve the exact copying of another animal's behavior, if only because the copier (particularly if a juvenile) may differ from the target in size, strength, and other relevant characteristics. The successful imitator must be able, as Tomasello and Call say, to separate the over all goal of the imitation from the particular means employed, viewing the latter as an independent step, and be able to copy the means at something more like a functional level – that is, in a way that reproduces those of its casual characteristics that are essential to produce the goal -- while at the same time varying other features to accommodate differences between the targets and imitators situation and abilities. This suggests (here I take myself to be following Tomasello and Call) that we should expect to find the following abilities occurring together: ability to imitate activities that have a means/ends structure, ability to learn about complex causal structures through combinations of interventions that reveal direct versus indirect causal relationships, ability to learn about causal relationships by observing the interventions of others, and a conception of causation according to which it is a tertiary relationship, and associated with this an ability to use information learned about causal relationships through passive observation to guide interventions and vice-versa. To a very substantial extent these abilities seem to be unique to humans, with non – human animals having abilities (a capacity for instrumental conditioning, ability to learn action

outcome sequences etc.) that have more of a stage one feel to them.

Since a number of the experiments that most clearly show that even small childen have these abilities have been performed by Gopnik, Schulz and others (Gopnik, Glymour, Sobel, Schulz, Kushir, and Danks, 2004, Gopnik and Schulz, 2004) and are described elsewhere in this volume, I will confine myself to a very brief overview, emphasizing general connections with the interventionist approach. First, young children learn not just the casual consequences of their own single interventions but, more interestingly, learn other causal relationships from *combinations* of interventions performed by others. They learn in conformity with a conditional intervention principle that is essentially just the definition of direct cause (**DC**). Moreover, they do this in contexts in which information about generative mechanisms, the transmission of force, and spatiotemporal clues cannot be used to identify the correct causal structure. For example, when confronted with a device with two interlocked gears, A and B that move together, which may also be influenced by the position of a switch, and which is such that the gears are removable, but only when the switch is off, the children are able to correctly infer that the motion of A causes B to move (when the switch is on) and that the motion of B does not cause A to move, not on the basis of intervening on A and observing the motion of B, but rather on the basis of information about what happens to A (B) when the switch is first turned off, B(A) is removed, and then the switch is turned on .In effect, this operation shows that the switch does not directly influence B without going through A while the switch influences A even with B fixed at the value "removed". Moreover, children can also acquire knowledge of causal structure from information about conditional probabilities and then use this information to predict the outcomes of new interventions or to produce new interventions that are appropriate for desired goals. That is, they can transfer or move back and forth between observational and intervention –based causal learning in a way that non – human animals apparently cannot.

The important role that learning from the interventions of others appears to play in the development of human casual understanding suggests that two abilities that are often regarded as rather different – the social cognition abilities involved in imitation and causal understanding of the non-social world -- may be closely intertwined. There is also independent evidence that young children are motivated to pay particular attention to the actions of other humans and that they have primitive imitative or simulative abilities for parsing and copying the actions of others humans. One might speculate that these attentional biases and abilities (which seem to be specific to humans in some respects) are combined with instrumental learning abilities that are shared with non-human animals to enable the much stronger forms of causal learning exhibited by humans.

References:

Bennett, J. (1984) "Counterfactuals and Temporal Direction" *Philosophical Review* 93: 57-91.

Bogen, J. 2004. "Analysing Causality: The Opposite of Counterfactual is Factual" *International Studiesin the Philosophy of Science 18.*

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

Dickinson, A. and Shanks, D. "Instrumental Action and Causal Representation". 1995 in Sperber, D., Premack, D, and Premack, A *Causal Cognition.* Oxford: Oxford University Press.

Dickinson, ,A. and Balleine, 2000 "Causal Cognition and Goal Directed Action" in C. Heyes and L. Huber, eds. *The Evolution of Cognition.* Cambridge, MA: MIT Press.

Glymour, C. 1998 " Learning causes: Psychological Explanations of Causal Explanation. *Minds and Machines* 8: 39-60.

Glymour, C. Forthcoming " Comment on D. Wegner, The Illustion of Conscious Will", *Behavioral and Brain Sciences.*

Gopnik, A. and Schulz, L. 2004 "Mechanisms of Theory Formation in Young Children" *Trends in Cognitive Science* 8: 371- 377.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L. Kushir, T. and Danks, D. "A Theory of Causal Learning in Children: Causal Maps and Bayes' Nets" *Psychological Review* 11: 3-22.

Harris, P. 2000 *The Work of the Imagination.* Oxford: Blackwell.

Hall, N. 2004, "Two Concepts of Causation" in Collins, J., Hall, N., and Paul, L. eds. *Causation and Counterfactuals* Cambridge, MA: MIT Press.

Hesslow, G. 1976 "Two Notes on the Probablistic Approach to Causality" *Philosophy of Science* 43: 290-92.

Hitchcock , C. 1995. "Discussion: Salmon on Explanatory Relevance" *Philosophy of Science* 62: 304-320.

Hitchcock, C. 2001 "The Intransitivity of Causation Revealed in Equations and Graphs" *Journal of Philosophy* 98: 273-99.

Kohler, W. 1927 *The Mentality of Apes* (2nd edition) New York: Vintage Books.

Lagnado, D. and Sloman, S. 2004 The Advantage of Timely Intervention" *Journal of Experimental Psychology: Learning, Memory and Cognition* 30:856-876.

Schaffer, J. 2000 "Causation by Disconnection" *Philosophy of Science* 67: 285-300.

Lewis. D. 1973 "Causation" *Journal of Philosophy* 70 556-67

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

Povinelli, D. 2000. *Folk Physics for Apes.* Oxford: Oxford University Press.

Premack, D. and Premack, A. *Original Intelligence*. New York: McGraw- Hill.

Salmon, W. 1994. "Causality Without Counterfactuals." *Philosophy of Science* 61: 297-312.

Schaffer, J. 2000 "Causation by Disconnection" *Philosophy of Science* 67: 285-300.

Schottmann, A. and Shanks, D. 1992, "Evidence for a Distinction Between Judged and Perceived Causality" *Quarterly Journal of Experimental Psychology* 44A: 321-42.

Schultz, T. 1982 Rules of Causal Attribution. *Monographs of the Society for Research in Child Development* 47.

Sloman, S. and Lagnado, D. Forthcoming. "Do We 'Do'? " *Cognitive Science*.

Sobel, D. and Kushnir, T. ,2003. "Interventions do not solely benefit causal learning: Being told what to do results in worse learning than doing it yourself". *Proceedings of the 2003 meeting of the Cognitive Science Society*, Boston, MA.

Sommerville, J. Forthcoming "Detecting structure in action: Infants as causal agents" This volume.

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

Steyvers, M. Tenenbaum, J. Wagenmakers, E. and Blum, B. (2003) "Inferring Causal Networks from Observations and Interventions*", Cognitive Science* 27, 453-89.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition*.

Visalberghi, E. and Trinca, L. (1989) "Tool Use in Capuchin Monkeys: Distinguishing Between Performance and Understanding" *Primates* 30: 511-21.

Wegner, D. (2002) *The Illusion of Conscious Will*. MIT Press: Cambridge, MA.

Woodward, J. 2002. "What is a Mechanism? A Counterfactual Account". In Jeffrey A. Barrett and J. McKenzie Alexander, *PSA 00*, Part II. *Philosophy of Science*, Supplement to Volume 69, No. 3, pp. S366-377.

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

See Hitchcock, 2001 and Woodward, 2003.

As Woodward, 2003, chapter 3 observes the arrow-breaking aspect of interventions reproduces a number of other features of Lewis' theory, including what look like "miracles".

For details, see Woodward, 2003, ch.5.

Although I lack the space for detailed discussion, it is worth observing that examples of this sort have implications for the claim, very common among both philosophers and psychologists, that interventionist counterfactuals hold only "in virtue" of facts about the existence of connecting causal processes and mechanisms, with these capturing what is really fundamental to causation. On its most staraightforward reading, this claim is simply false: from the fact that there is a connecting causal process, transmission of energy etc. from the motion of the cue stick to trajectory of the eight ball, we can deduce almost nothing about which interventionist counterfactuals associated with this process are true. Any explanation of why these interventionist counterfactuals hold will need to appeal to generalizations that are far more specific (eg. the laws of conservation of energy and momentum) and it is plausible that these will also have a counterfactual element built into them.

It is thus a mistake to think of a plausible account of mechanisms and an interventionist account of causation (or Bayes' net approaches) as in opposition to one another. See Glymour, 1998 for a similar view.

In the interests of moving the discussion along, I am riding roughshod over a number of complications and possibilities. Of course it is possible to hold that there is no straightforward correspondence between what causation is and how we think about it—this was Hume's view on one natural interpretation. My view is that positions of this sort are not interesting when advanced as mere logical possibilities – instead the way in which people think and learn about causation (and how these fail to correspond to what causation is) needs to be spelled and it needs to be shown how these explain known experimental results. Unlike Hume, contemporary philosophers rarely do this.

As Gopnik has noted, the obvious analogue here is with vision. People don't just have visual experiences and make visual judgments – in addition these are often veridical. An adequate theory should explain how this happens.

In contrast, there is evidence (Schottmann and Shanks, 1992) that causal *perception* of collision phenomena is *not* sensitive to such contingency information, although *judgment* of causal efficacy is. In other words process theories fit better with causal perception than causal judgment tasks. One may thus conjecture that causal perception phenomena explain some of the intuitions that underlie causal process theories.

To guard against possible misunderstanding, let me say explicitly that I do not regard it as a *necessary* condition for subjects to possess and to be guided by an interventionist conception of causation that they be able to reason explicitly with counterfactuals: subjects also can possess an implicit understanding of aspects of counterfactual reasoning, as revealed, for example, in non-verbalized planning. The argument above is simply that explicit use of counterfactual reasoning and explicit recognition of its connection to causal claims is *sufficient* to establish that subjects are operating with a broadly counterfactual conception of causation. A similar point holds for subjects explicit recognition of the

connection between causal claims and interventions.

Exactly why this is true is a matter of on-going discussion – see Sommerville, this volume, Lagnado and Sloman, 2004 for some alternative suggestions.

For some suggestions among broadly similar lines, see Glymour forthcoming.

See Wegner, 2002. Of course, as Wegner documents, there are illusions of agency but their existence does not show (and Wegner does not claim) that the feeling of agency is generally an unreliable clue to voluntariness.

An example: very shortly after inserting the key to unlock my car door, a car alarm goes off in a neighboring car, leaving me with the very strong impression that my action has caused the alarm to go off

Whether this is correct is of course an empirical matter. I don't claim that it obvious, merely that it is a conjecture which is worth exploring.

Thanks to Daniel Povinelli for a very helpful conversation that corrected a serious misunderstanding of his views in a previous draft.

Indeed, it might be argued, uncharitably, that the apes behave pretty much as though they *are* guided by this heuristic and that this simply shows what a gap there is between use of the heuristic and full-fledged causal understanding. In this connection, it is also worth noting that if the human possession of the concept of "force" is closely linkedto the abilities displayed in launching experiments, as both Povinelli and Tomasello and Call suggest and if apes fail to possess such a concept, it would seem to follow that they will behave quite differently from human children in, e. g,. looking time experiments involving launching phenomena. My prediction is that there will be no dramatic difference, again illustrating that, when linked to launching phenomena in the way described, possession of the concept of force is not sufficient for the kind of causal understanding displayed by humans.

As the passage quoted above makes clear, the experiments it describes have not actually been performed. It would be very worthwhile to do them.

David Danks, personal communication; Alison Gopnik, personal communication. Needless to say, it would be very worthwhile to explore this issue in the context of primate causal understanding by means of more systematic experiments.

Note also that there is nothing "unobservable" about this intermediate variable – that is, if the apes fail at the task under discussion, it is not because they fail to postulate *unobservable* intermediate variables but rather because they fail to recognize the relevance of an observable intermediate variable.

The systematic interrelationships between causal understanding and the ability to discern one's own intentions and goals as well as those of others is also one of the main themes of Sommerville's essay in this volume.

See Gopnik and Schulz, 2004, for a similar line of thought.

PAGE

PAGE 1