Reasons as Causes in Bayesian Epistemology

Clark Glymour and David Danks

Carnegie Mellon University and Florida Institute for Human and Machine Cognition

In everyday matters, as well as in law, we allow that someone's reasons can be causes of her actions, and often are. That correct reasoning accords with Bayesian principles is now so widely held in philosophy, psychology, computer science and elsewhere that the contrary is beginning to seem obtuse, or at best quaint. And that rational agents should learn about the world from energies striking sensory inputs—nerves in people—seems beyond question. Even rats seem to recognize the difference between correlation and causation,[i] and accordingly make different inferences from passive observation than from interventions. A few statisticians aside,[ii] so do most of us. To square these views with the demands of computability, increasing numbers of psychologists and others have embraced a particular formalization, causal Bayes nets, as an account of human reasoning about and to causal connections.[iii] Such structures can be used by rational agents, including humans in so far as they are rational, to have degrees of belief in various conceptual contents, which they use to reason to expectations, which are realized or defeated by sensory inputs, which cause them to change their degrees of belief in other contents in accord with Bayes Rule, or some generalization of it. How is all of this supposed to be carried out?

1. Representing Causal Structures

The causal Bayes net framework adopted by a growing number of psychologists goes like this: Our *representations* of causal relations are captured in a graphical causal

model, or causal Bayes net. We reason implicitly as though we were calculating explicitly (but often not quite accurately) with such a network in hand. The network is a mathematical object describing relations among features of a system or situation that are potentially variable—for example, having at least *present* or *absent* as possible values. Those features are vertices, or variables, in a network with directed edges from some vertices to others. A set of conditional probabilities is associated with the network, specifying for each vertex, *V*, the probability of each of its values conditional on each specification of values of the vertices in the graph that are parents of *V*—i.e., those that have edges directed into *V*. The graph is almost always assumed to be acyclic: there is no sequence of directed edges leading from a variable back to that same variable. For example, a simple network relating a lamp to an electrical power source and a switch on a timer might be:

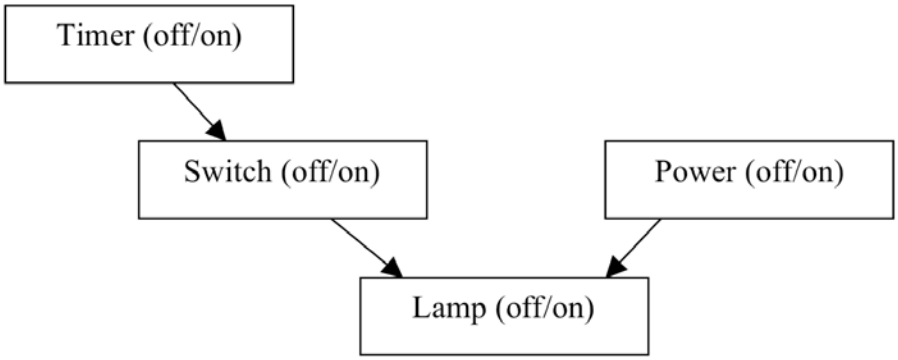

Figure 1: The Causal Bayes Net Ascribed to the World

Power and Switch have independent probability distributions. The state of Lamp is determined uniquely by its two inputs: Lamp is on if Power is on and Switch is on. If the value of Power is ignored or unknown and varies from case to case, then the state of Lamp will appear to be an indeterministic function of Switch. The causal content is captured by the supposition that a direct intervention that changes the state of a variable

changes the state of variables downstream from it, but leaves the state of other variables unchanged.[iv] So, for example, an intervention that breaks the bulb of the Lamp leaves the state of the Lamp fixed at off, regardless of any variation in Power, Switch or Timer, and leaves the probabilities of Power, Switch and Timer unaltered, as if the edges from Switch and Power to Lamp were broken by the intervention, though the arrow from Timer to Switch is left unaltered.

The causal attribution carries with it expectations—in the human, not the probabilistic, sense—about the joint frequencies of events, and also expectations about the results of possible or hypothetical interventions on features of the system. If an agent believes a light switch causes a lamp to go on, then the agent expects that turning the switch on and off will turn the lamplight on and off. The agent can use the Bayes net to reason to a degree of belief in a particular event, or even to a specific predicted value for a variable, if other variables in the network have specified values: if the light switch is believed to be chancy (but not overly so), the agent may derive a degree of belief that the lamp goes on when the switch is thrown, or may conclude that the lamp will go on. In the other direction, the agent can use the Bayes net to reason backwards to alter her degrees of belief about the possible causes of some observed event, say that the lamp is on, in accord with Bayes Rule. If the agent's prior degree of belief that the switch is on is r, that the lamp light is on is p, and that the lamp light is on given that the switch is thrown is q, then after coming to believe that the lamp light is on, the agent's degree of belief that the switch is on changes to rq/p. Further, the agent can use the Bayes net to reason hypothetically about the results of possible interventions that would fix one or more features from outside the system, and in particular, about the results of her or others'

actions. Suppose the timer is set to turn on the light switch automatically at a certain time. The effect on the state of the lamp of an outside intervention that turns *off* the light switch at some later time is reasoned about hypothetically by supposing that the value of Switch is fixed in the Causal Bayes Net Ascribed to the World, but nothing else is altered. In particular, the degrees of belief in the Timer state and that the lamp is on given that the light switch is off are left the same. Switch, which formerly depended on its parent variable in the graph (Timer), now becomes independent (in the degree of belief measure) of its parent. Graphically, the intervention breaks the directed edges from the parent variables—whatever they may be—to Switch. The results are expected to match the causal consequences of the corresponding intervention in the world.

This nice theoretical picture is substantiated by a variety of experiments that suggest that even young children make predictions and provide explanations that are patterned as a Bayes net requires, and change their confidence in outcomes roughly according to Bayes Rule. Some bits of the account are better established than others, and of course people make errors and are computationally limited. For example, no account is given of how people choose to attend to one phenomenon rather than another, and in psychological experiments that focus is almost always provided by the experimenter: *Rector ex machina*. Computer science aids the psychologists' account by providing a variety of algorithms for (comparatively) efficiently computing conditional probabilities in a Bayes net, and for computing probabilities given an intervention; that is, regardless of whether people make inferences just as the computer algorithms do, the inferences are at least feasible. And, finally, recent work has shown that neural firing frequencies in a recurrent neural network—one with feedback loops—can implement an algorithm that

computes some of the conditional probabilities defined in a Bayes net. Moreover, the neural model corresponds well with firing frequencies observed in the visual cortex.[v]

But can reasons like these, observations of features of the world that are causes or effects of other features, be causes? We do not mean to suggest somehow that the degrees of belief, or changes in them, are epiphenomenal and therefore not causal simply because the computations of conditional probabilities are carried out by neural processes; we are content with local identifications of changes in degrees of belief with instances of neural processes. Our concern is rather with how and whether the reasoning that psychologists suppose agents do with The Causal Bayes Net Ascribed to the World can itself be consistently represented using a causal Bayes net, as should be possible if those reasons are causes.

2. Connecting Causal Beliefs and Inference

Let us assume (for the moment) that the connections and mechanisms needed for computing probabilities according to the Timer → Switch → Lamp ← Power network are somehow implemented in a reasoning agent. Suppose now the agent wishes the lamp to light at 6:00. Her reasoning to a timer setting presumably goes something like this: "If I set the timer for 6:00, then the switch will go on at 6:00. If the power is on at 6, then the lamp will certainly light at 6. The timer setting is independent of whether the power is on at 6. It is very probable that the power will be on at 6. Therefore, if I set the timer for 6, then the light will very probably go on at 6."

So she sets the timer to go on at 6:00, and expects the lamp to go on at 6:00. Her reasons include both a desire and a sequence of degrees of belief about consequences of an action. The reasons are causes, not only of her action, but also of the change in her

degrees of belief that the switch will go on at 6 and that the lamp will go on at 6. As causes, her degree of belief reasons mirror the structure of the causal Bayes net structure she ascribes to the Timer/Switch/Lamp/Power system, but the variables are now her own degrees of belief in various conceptual contents. The goal that the light go on at 6, whether hypothetical or desired, somehow determines the relevant variables for the Causal Bayes Net Ascribed to the World (since there must be a great many such causal networks available to the agent), and the course of reasoning to the conditional forecast:

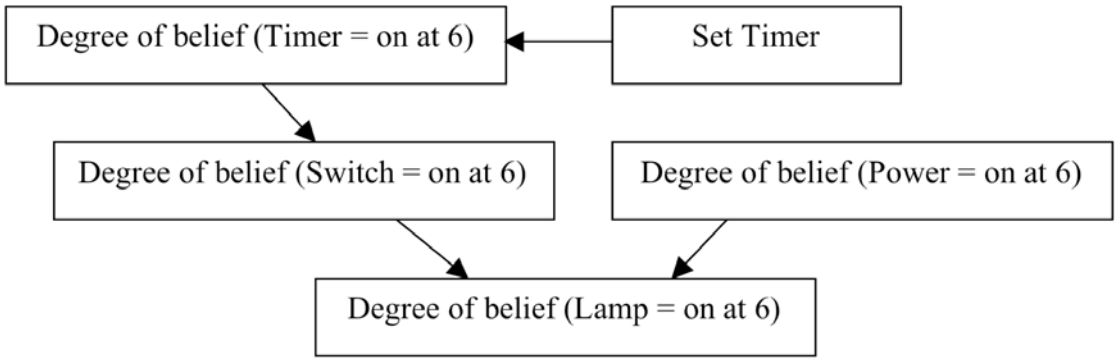

Figure 2: The Causal Bayes Net of Reasoning to a Forecast

The variables now range over values of degrees of belief—whether these are all real values between 0 and 1, or some finite range, as in high, medium, low, makes no difference here.

The relations between degrees of belief in The Causal Bayes Net of Reasoning to a Forecast are chancy, as they describe some causal process in the brain, which may be subject to various chance fluctuations. Accordingly, there are conditional probabilities associated with the directed graph, and in agreement with the psychological hypothesis that causes are represented as a graphical model, we will assume these conditional probabilities together determine a joint probability distribution.[vi]

Suppose now that the lamp does not light at 6:00, and the agent sees that it does not light. According to the psychological story, she should then reason using the Causal Bayes Net Ascribed to the World by conditioning on Lamp = off to compute a new probability that the power is off, and a new probability that the Switch and/or Timer are off. These new computations will constitute reasoning from the observations to new degrees of belief corresponding to new probability ascriptions; the observation about the lamp is a reason to change one's belief about other features of the world. That picture seems eminently reasonable; any serious epistemological theory holds that we use observations to change our degrees or strengths of belief. But that picture does not work with The Causal Bayes Net of Reasoning to a Forecast just described.

The Causal Bayes Net of Reasoning to a Forecast specifies, prior to the agent perceiving at 6 that the Lamp is off, the causes of the agent's degree of belief that the Lamp is (or will be) on at 6. Those causes are her prior degrees of belief that the Switch is on at 6 and her prior degree of belief that the Power is on at 6, and more remotely, her prior degree of belief that the Timer is on at 6. In other words, perception of the light state is not a cause of degree of belief in the light state *from the point of view of this system*. Thus, the perception that the Lamp is off is an intervention on her Degree of belief that the Lamp is on, and so the perception that the Lamp is off at 6 cannot alter any of her degrees of belief in the other propositions, exactly because it is an intervention on Degree of belief(Lamp = on) in The Causal Bayes Net of Reasoning to a Forecast. More formally, all of the edges into Degree of belief(Lamp = on) are broken, and so each pathway from other variables to Degree of belief(Lamp = on) is destroyed by the intervention. Therefore, by the Markov property, all of the other Degree of belief

variables are independent of Degree of belief(Lamp = on) after the perception. But by Bayes Rule, a proposition that is independent of a piece of evidence is not changed by acquiring the evidence, and the values of other nodes in the Causal Bayes Net of Reasoning to a Forecast are therefore unaltered by the perception of the lamp state.

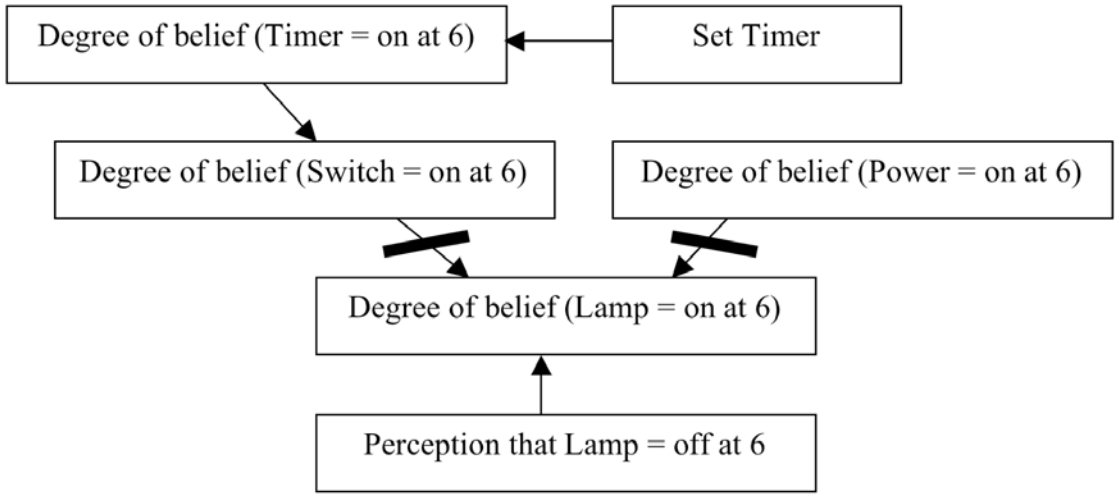

Figure 3: The Causal Bayes Net of Reasoning to a Forecast after Perception

The psychological story has a problem: the view that degrees of belief, or changes in them, are causes seems incompatible with Bayesian learning from perception. Perception of the state of an effect should lead (by Bayesian updating) to changes in beliefs about the causes, but perception is an exogenous intervention in the standard reasoning network, and so breaks the connections between the effect and its causes.

Qualitatively, the agent's reasoning upon perceiving that the lamp is not lit at 6 goes something like this: "The lamp is not on, therefore the probability that the power is on is decreased and the probability that the switch is on is decreased; because the probability that the switch is on has decreased, the probability that the timer is on is also decreased." The unmentioned sensation is the initial cause of this sequence of reasons. If reasons are causes in this particular type of reasoning from perception to new degrees of

belief in other contents, then the internal causal chain starting with the new degree of belief that the Lamp is on goes like this:

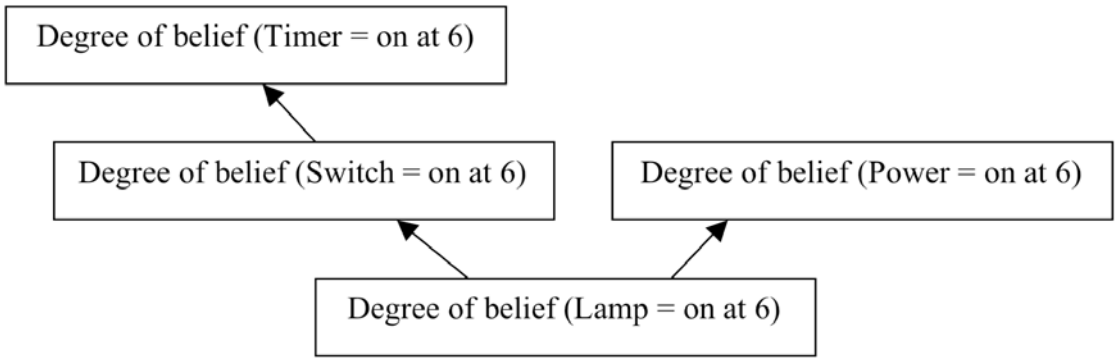

Figure 4: The Causal Bayes Net of Reasoning from Perception

If reasons are causes, then the causal structure for reasoning from perception is exactly reversed from the causal structure for reasoning to a forecast. If Degree of belief(X) → Degree of belief(Y) in The Causal Bayes Net of Reasoning to a Forecast, then Degree of belief(X) ← Degree of belief(Y) in The Causal Bayes Net of Reasoning from Perception. The reversal of edges, however, means that the two causal Bayes nets—one of forecasting and the other of perception—do not agree in the constraints they impose on the joint probability distribution for degrees of belief. In this example, the Markov property applied to the Causal Bayes net of Reasoning to a Forecast, implies that Degree of belief(Power = on) = x is unconditionally independent of Degree of belief(Switch = on) = y, for all x, y. The Causal Bayes Net of Reasoning from a Perception implies no such thing. We are whipsawed.

3. Combining the Bayes Nets

Human perception, we think, is often in part top-down, driven by prior conceptual structure and prior degrees of belief. For a Bayesian agent whose reasons are causes, the problems just discussed suggest that perception that accords or conflicts with a prior

degree of belief should have a top-down contribution. In order for sensation to cause our imagined rational agent to form a new degree of belief that the lamp is lit, and to do so in a way that allows a Bayesian updating of the value of the agent's degrees of belief that the switch is on and that the power is on, the new degree of belief that the lamp is lit must be the collaborative, interactive effect of the values of her degree of belief that the switch is on, of her degree of belief that the power is on, and of the sensory input. The sensor input does not *itself* change the value of Degree of belief(Lamp = on), but rather it changes the degree of belief that the lamp is lit *given the values of parents of Degree of belief(Lamp = on)* in the causal Bayes Net of Reasoning to a Forecast. For different values of Degree of belief(Power = on) and Degree of belief(Switch = on), the input of sensation will result in different values of Degree of belief(Lamp = on), and so the intervention of sensation will not make the agent's Degree of belief that the Lamp is on independent of her other Degrees of belief.

Since reasoning goes from beliefs about circumstances to forecasts of perceptions, and from perceptual changes in belief to new beliefs about circumstances, it seems that the "reasons are causes" view requires a representation of the causal connections that likewise goes in both directions. It seems that we need, in other words, a cyclic causal graph among degrees of belief, with appropriate associated probabilities.

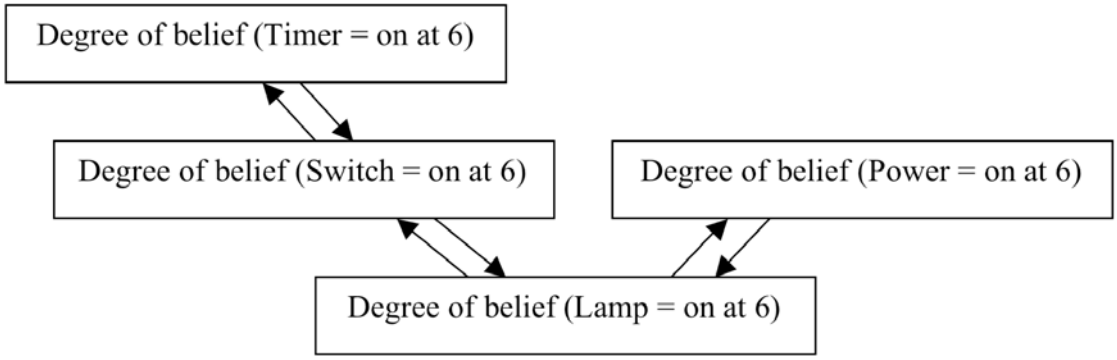

Figure 5: The Combined Causal Bayes Net

The problem is we do not know much about cyclic graphical representations of causal relations, or how to update them by Bayes rule, and what we do know is problematic for this view. In the scientific literature, causal Bayes nets are generally taken to be acyclic, but that is not strictly necessary. One can have networks with cycles, even with cycles that have edges in each direction between two variables. Probabilistic constraints that generalize the Markov property for acyclic networks still hold, necessarily, for linear cyclic systems, and can consistently be assumed for cyclic networks with variables that have a finite range of possible values. So we might consider whether the sensory input can nudge the degrees of belief in values of a variable, which nudges the degrees of belief in its parents, which nudges the degrees of belief in the variable again, which nudges… and so on, until an equilibrium is reached.

That is certainly possible, but there are two related difficulties: How can updating on evidence occur, and can it be Bayesian? Consider the second difficulty first, in the simplest case in which the variable that is directly influenced by sensory inputs, denote it by *S*, has a single parent variable, *Y*. The idea is that the value of *S* causes the value of *Y* to be updated, which causes the value of *S* to be updated, and so on, until no more changes result. On the Bayesian perspective, each step in each direction, no matter how implemented in the brain, should result in updating one of the variables conditional on the currently updated value of the other, and we should therefore expect that at equilibrium the joint degree of belief in *S* and *Y* together should be the product of their conditional probabilities on each other: for all values of *S* and *Y*, DOB(*S*,*Y*) = DOB(*S* | *Y*)DOB(*Y*| *S*). But this equation implies that *Y* and *S* are independent![vii] Applied to our

example, upon learning that the lamp is not lit at 6, the agent's degrees of belief would then be altered in such a way that the degree of belief that the switch is on and the degree of belief that the timer is on have no relation to one another. We should not welcome such a theory.

Not only does a Bayes Rule requirement for updating lead to absurd results in cyclic networks, no correct updating algorithm is known for such systems and certainly no algorithm of the kind that neural systems plausibly implement for acyclic Bayes nets.[viii] Some other resolution is needed.

4. Dynamics to the Rescue?

The general problem is that the causal direction of influences of degrees of belief must go one way when forecasting, and the reverse direction when learning from experience, and the conditional probabilities of the changes must be in phase. If we can presume that forecasting and learning are not simultaneous, then there is a Bayesian solution, using structures that are sometimes called "dynamical Bayes nets" but which are really the same sorts of structures we have considered so far, except that the variables—in this case degrees of belief—are indexed by time. Consider the structure:

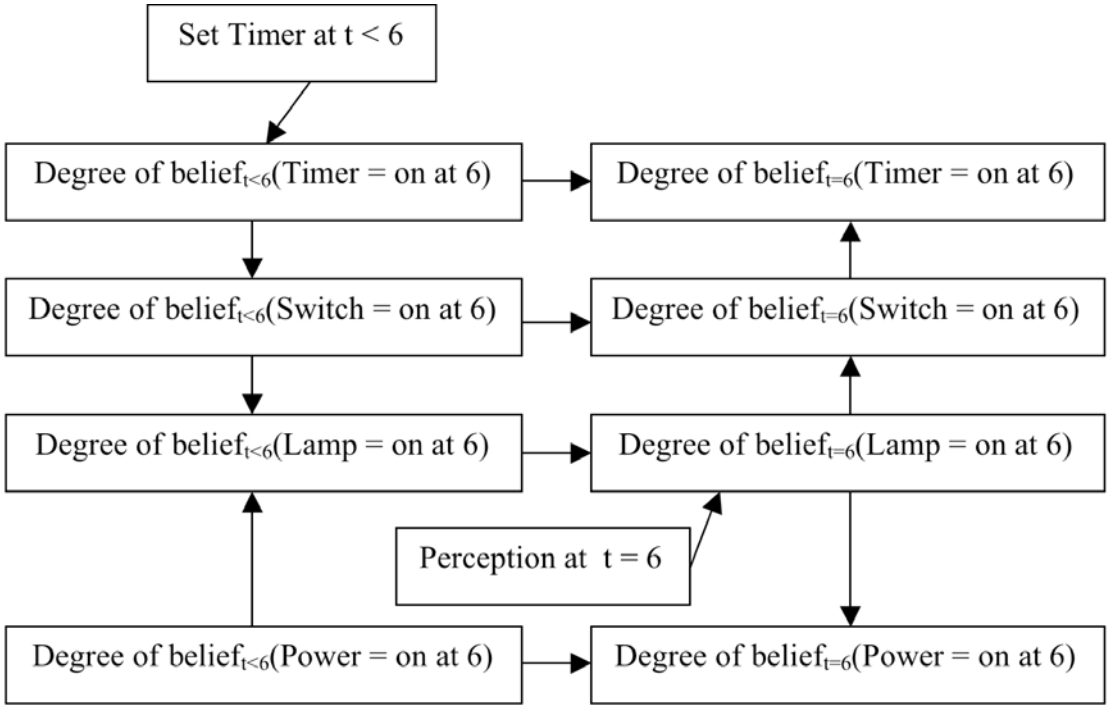

Figure 6: The Dynamical Bayes Net of Reasoning to Forecasts and from Perception

All of the probabilities of conditional degrees of belief in this network can be consistently estimated by Bayes Rule, even while the values of the variables—the degrees of belief—are themselves determined, up to chance variation, by Bayes Rule applied to the external evidence, setting the timer and sensation.[ix]

So we have a solution in which reasoning is Bayesian almost all the way through, and reasons are causes. We do not know of any neural implementation of dynamic Bayes nets, and any neural realization that involves both forecasting and learning, rather than only visual recognition, will not be localized in the visual cortex. Verifying the hypothesis that, in humans, reasoning is Bayesian and reasons are causes would be difficult, but at least we have some idea of the conditions that need to be satisfied.

However, this is in some respects an odd and incomplete solution. As in our story, agents must be able to reason hypothetically, for otherwise we could not rationally plan.

But hypothetical reasoning requires hypothetical degrees of belief: the expectations one would have if one believed some set of assumptions, be they assumptions about causal structure or assumptions about values of variables ascribed, hypothetically, to the world, or assumptions of some other kinds. When reasoning under a hypothesis, hypothetical degrees of belief, or their changes, ought still to be causes of other hypothetical degrees of belief in something like the way we have described, at least if the Bayesian, "reasons are causes," account is correct. There is no difficulty in this from the point of view of an omnipotent Bayesian calculator (the kind philosophy generally assumes), and there is no difficulty in principle from the point of view of a programming system, provided the number of alternative hypotheses is not too large: the degrees of belief on each hypothesis are computed, and averaged with weighting by the degrees of belief in the various hypotheses. But how a neural system could implement such reasoning, distinguishing between the hypothetical and the all-things-considered non-hypothetical degrees of belief, remains to be discovered…or not. And that is not the only inadequacy.

The agent must have just the right course of reasoning instantiating just the right Dynamic Bayes Net, and that must somehow, mysteriously, be determined by the agent's goals and epistemic circumstances at the moment. For example, if the agent is not home at 6 p.m. to perceive the state of the lamp, she may follow the forecast to 6 p.m. with another forecast from 6 p.m. If instead of seeing the state of the lamp, she sees the power is off, she will have a quite different sequence of changes of belief. The correct course of reasoning must somehow be generated on the fly, and we have no account of how that is done. The machine still has its ghost.

[i] Aaron P. Blaisdell, Kosuke Sawa, Kenneth J. Leising, and Michael R. Waldmann, "Causal Reasoning in Rats," *Science*, CCCXI (2006): 1020-22.

[ii] E.g., Karl Pearson, *The Grammar of Science* (Mineola: Dover, 2004). Still in print, Pearson maintains that causation is correlation and that there is no fact to the matter of correlation or causation because what we experience is events in our brains.

[iii] A very limited sample, each with references to other relevant works, includes the following: Clark Glymour, *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology* (Cambridge: MIT, 2002); Alison Gopnik, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks, "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets," *Psychological Review*, CXI (2004): 3-32; Alison Gopnik and Laura E. Schulz, eds., *Causal Learning: Psychology, Philosophy, and Computation* (Oxford: Oxford, 2007); Steven A. Sloman, *Causal Models: How People Think about the World and Its Alternatives* (Oxford: Oxford, 2005).

[iv] A framework for specifying the consequences of interventions in a network is given in Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, Prediction, & Search* (Berlin: Springer-Verlag, 1993). Algorithms for computing the effects of interventions that fix a definite value for one or more variables are given in Judea Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge, 2000), and discussed as a foundation for understanding causation by James Woodward, *Making Things Happen: A Theory of Causal Explanation*, (Oxford: Oxford, 2003).

[v] Rajesh R.N. Rao, "Bayesian Inference and Attentional Modulation in the Visual Cortex," *Cognitive Neuroscience and Neuropsychology*, XVI (2005): 1843-48. See also:

Kenji Doya, Shin Ishii, Alexandre Pouget, and Rajesh P. N. Rao, eds., *Bayesian Brain: Probabilistic Approaches to Neural Coding*, (Cambridge: MIT, 2007).

[vi] We assume the chances satisfy the Markov property for the graph: the joint probability distribution is the product of the conditional distribution of each variable in the graph given values of all its parent variables.

[vii] $Pr(S \,\&\, Y) = Pr(S \mid Y)Pr(Y \mid S) = Pr(S \,\&\, Y)Pr(Y \,\&\, S) / Pr(Y)Pr(S) \Rightarrow Pr(Y)Pr(S) = Pr(Y \,\&\, S)$.

[viii] However, procedures for computing correlations resulting from interventions on a single variable in linear cyclic systems have been known for a long time.

[ix] Strictly, this implements the representation of interventions proposed in Spirtes, et al., *op. cit.* rather than the representation in Pearl, *op. cit.*