Learning Causal Relationships

Jon Williamson*

presented at
'Causality: Metaphysics and Methods'
LSE 20-21 June 2002

Abstract

How ought we learn causal relationships? While Popper advocated a hypothetico-deductive logic of causal discovery, inductive accounts are currently in vogue. Many inductive approaches depend on the causal Markov condition as a fundamental assumption. This condition, I maintain, is not universally valid, though it is justifiable as a default assumption. In which case the results of the inductive causal learning procedure must be tested before they can be accepted. This yields a synthesis of the hypothetico-deductive and inductive accounts, which forms the focus of this paper. I discuss the justification of this synthesis and draw an analogy between objective Bayesianism and the account of causal learning presented here.

Contents

*Department of Philosophy, King's College, London. jon.williamson@kcl.ac.uk.

How ought we learn causal relationships? While there has long been abundant philosophical interest in the metaphysics of causality, there has until recently been relatively little discussion of this fundamental epistemological question. Popper addressed the issue in his 'Logic of Scientific Discovery',[1] and recently there has been a spate of inductive accounts and controversy surrounding them. I shall present the Popperian and the inductive approaches, and their key shortcomings, in §1 and §2 respectively. I will argue in §3 and §4 that a synthesis of these two approaches overcomes these shortcomings. This answer to the epistemological question is compatible with various metaphysical positions on the nature of causality, as indicated in §5.

§1
Hypothetico-Deductive Learning

Karl Popper's account of causal discovery is derived from his general picture of scientific discovery. For Popper a causal explanation of an event consists of natural laws (universal statements) together with initial conditions (singular statements) from which one can predict (by deduction) the event to be explained. The initial conditions are called the 'cause' of the event to be explained, which is in turn called the 'effect'.[2] Causal laws, then, are just universal laws, and are to be discovered via the familiar Popperian scheme: (i) hypothesise the laws; (ii) deduce their consequences, rejecting the laws and returning to step (i) if these consequences are falsified by evidence. Popper thus combines a 'covering-law' account of causal explanation with a hypothetico-deductive account of learning causal causal relationships.

The covering-law model of explanation was developed by Hempel and Oppenheim[3] and Railton,[4] and criticised by Lewis.[5] While such a model fits well with Popper's general account of scientific discovery, neither the details nor the viability of the covering-law model are relevant to the issue at stake: a Popperian hypothetico-deductive account of causal discovery can be combined with practically any account of causality and causal explanation.[6] Neither does one have to be a strict falsificationist to adopt a hypothetico-

[1] [Popper 1934].
[2] [Popper 1934] §12.
[3] [Hempel & Oppenheim 1948].
[4] [Railton 1978].
[5] [Lewis 1986] §VII.
[6] Even Russell's eliminativist position of [Russell 1913] ties in well with Popper's logic of scientific discovery. Popper, after all, drew no sharp distinction between causal laws and the other universal laws that feature in science.

deductive account. Given a suitable notion of confirmation one can accept the hypothesised causal relationships according to the extent to which evidence confirms deduced predictions.

Besides providing some criterion for accepting or rejecting hypothesised causal relationships, the proponent of a hypothetico-deductive account must do two things: (i) say how causal relationships are to be hypothesised; (ii) say how predictions are to be deduced from the causal relationships.

Popper fulfilled the latter task straightforwardly: effects are predicted as logical consequences of laws given causes (initial conditions). The viability of this response hinges very closely on Popper's account of causal explanation, and the response is ultimately inadequate for the simple reason that no one accepts the covering-law model as Popper formulated it: more recent covering-law models are significantly more complex, coping with chance explanations.[7]

Popper's response to the former task was equally straightforward, but perhaps even less satisfying:

> my view of the matter, for what it is worth, is that there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process. My view may be expressed by saying that every discovery contains 'an irrational element', or 'a creative intuition'[8]

Popper accordingly placed the question of discovery firmly in the hands of psychologists, and concentrated solely on the question of the justification of a hypothesis.

The difficulty here is that while hypothesising may contain an irrational element, Popper has failed to shed any light on the rational element, which is surely significant. Popper's scepticism about the existence of a logic need not have precluded him from discussing the act of hypothesis from a normative point of view: both Popper in science and Pólya in mathematics remained pessimistic about the existence of a precise logic for hypothesising, yet Pólya managed to identify several imprecise but important heuristics.[9] One particular problem is this: a theory may be refuted by one experiment but perform well in many others; in such a case it may need only some local revision, to deal with the domain of application on which it is refuted, rather than wholesale rehypothesising. Popper's account says nothing of this, giving the impression that with each refutation one must return to a blank sheet

[7] [Railton 1978] for example.

[8] [Popper 1934] pg. 32.

[9] [Pólya 1945], [Polya 1954], [Polya 1954b].

and hypothesise afresh. The hypothetico-deductive method as stated neither gives an account of the progress of scientific theories in general, nor of causal theories in particular.

Any hypothetico-deductive account of causal discovery which fails to probe either the hypothetico or the deductive aspects of the process is clearly lacking. These are, in my view, the key shortcomings of Popper's position. I shall try to shed some light on these aspects when I present a new type of hypothetico-deductive account in §3. For now, we shall turn to Popper's arch-enemy, inductivism.

§2
Inductive Learning

Inductive learning involves learning causal relationships directly from data. Usually, but not necessarily, the data is statistical and the relationships are general-case ('type-level') rather than single-case ('token-level'). Such an account of learning is occasionally alluded to in connection with probabilistic analyses of causality and has been systematically investigated by researchers in the field of artificial intelligence (AI), including groups in Pittsburgh,[10] Los Angeles[11] and Monash,[12] proponents of a Bayesian learning approach,[13] and computationally-minded psychologists.[14]

The AI approaches seek to learn various types of causal model. A *causal graph* is a purely qualitative model — this is a directed acyclic graph in which nodes correspond to variables and there is an arrow from one node to another if the former directly causes the latter. A *probabilistic causal model* also contains quantitative information pertaining to the strengths of the causal connections: it consists of a causal graph together with the probability distribution of each variable conditional on its direct causes. For each causal link $C \longrightarrow E$, the probability distribution of E conditional on C and E's other causes $D_1, \ldots, D_k$ is provided; this can be used to determine the degree to which changing the value of C from c to c' brings about the

[10] [Spirtes et al. 1993], [Scheines 1997], [Glymour 1997], [Mani & Cooper 1999], [Mani & Cooper 2000], [Mani & Cooper 2001].

[11] [Pearl 2000], [Pearl 1999].

[12] [Dai et al. 1997], [Wallace & Korb 1999].

[13] [Heckerman et al. 1999], [Cooper 1999], [Cooper 2000], [Tong & Koller 2001], [Yoo et al. 2002].

[14] [Waldmann & Martignon 1998], [Waldmann 2001], [Tenenbaum & Griffiths 2001], [Glymour 2001], [Hagmayer & Waldmann 2002], www-psych.stanford.edu/ jbt/causal-workshop.html

value e of E, via the expectation $\mathcal{E}_{D_1,\ldots,D_k} p(e|c'D_1 \ldots D_k) - p(e|cD_1 \ldots D_k) = \sum_{d_1,\ldots,d_k} p(d_1 \ldots d_k)[p(e|c'd_1 \ldots d_k) - p(e|cd_1 \ldots d_k)]$. A third type of causal model is a *structural equation model*. This can be thought of as a causal graph together with an equation for each variable in terms of its direct cause variables, $E = f(C, D_1, \ldots, D_k, \varepsilon)$, where f is some function and ε is an error variable. Approaches differ as to whether they seek to learn a single 'best' causal model given the data, or a class of causal models, each of which is consistent with the data.

The mainstream of AI approaches have the following feature in common. In order that causal relationships can be gleaned statistically, the approaches assume a strong connection between causality and probability, namely the causal Markov condition.[15] This assumes that each variable in the causal graph is probabilistically independent of its non-effects, conditional on its direct causes.[16] The inductive procedure then consists in finding the class of causal graphs — or under some approaches the 'best' causal graph — whose probabilistic independencies implied via the causal Markov condition are consistent with independencies inferred from the data. Other assumptions are often also made, such as minimality (no subgraph of the causal graph also satisfies the causal Markov condition), faithfulness (all independencies in the data are implied via the causal Markov condition), linearity (all variables are linear functions of their direct causes and uncorrelated error variables), causal sufficiency (all common causes of measured variables are measured), context generality (every individual possesses the causal relations of the population), no side effects (one can intervene to fix the value of a variable without changing the value of any non-effects of the variable) and determinism. However these extra assumptions are less central than the causal Markov condition: approaches differ as to which of these extra assumptions they adopt and the assumptions tend to be used just to facilitate the inductive procedure based on the causal Markov condition, either by helping to provide some justification for the inductive procedure or by increasing the efficiency or efficacy of algorithms for causal induction.

There are a number of possible difficulties with the inductive approach.

[15]There are inductive AI methods that take a totally different approach to causal learning, such as that in [Karimi & Hamilton 2000] and [Karimi & Hamilton 2001], and [Wendelken & Shastri 2000]. However, non-causal-Markov approaches are well in the minority.

[16]A probabilistic causal model under the assumption of the causal Markov condition is called a *Bayesian network*. In the case of structural equation models the causal Markov condition is a consequence of the representation of each variable as a function just of its direct causes and an error variable, given the further assumption that all error variables are probabilistically independent.

One might express conceptual reservations about taking variables as relata of causality and taking the domain-dependent relation of direct causality as primitive. Alternatively, the prerequisites such as the causal Markov condition, acyclicity of the causal graph and the extra assumptions might bear the brunt of criticism. The proof, inductivists claim, will be in the pudding. However, the reported successes of inductive methods have been questioned,[17] and these criticisms reflect on the approach as a whole and the causal Markov condition in particular as its central assumption. Indeed, the causal Markov condition has received a great deal of criticism with a number of compelling counterexamples put forward in the literature. These causal Markov counterexamples are well known and are catalogued and discussed in some detail in §2 of [Williamson 2001] — I shall assume their cogency in this paper without further ado.[18]

The key shortcoming of the inductive approach is, in my view, this: given the causal Markov counterexamples, the inductive approach cannot guarantee that the induced causal model or class of causal models will tally with causality as we understand it. The causal models that result from the inductive approach will satisfy the causal Markov condition, but the real causal picture may not.[19] I shall argue next that while this may put paid to the dream of using causal Markov formalisms for learning causal relationships via a purely inductive method, the formalisms should not themselves be abandoned because they feature in a modified account of causal learning.

§3
A Synthesis

The approach I would like to put forward proceeds along the following lines.

Induce AI learning techniques are used to hypothesise a causal model from data;

Predict predictions are deduced from the hypothesised model;

[17] [Humphreys & Freedman 1996], [Humphreys 1997], [Freedman & Humphreys 1999], [Woodward 1997].

[18] See also [Dash & Druzdzel 1999], [Hausman 1999], [Hausman & Woodward 1999], Part Three of [Glymour & Cooper 1999], [Lemmer 1996], [Lad 1999], [Cartwright 1997], [Cartwright 1999] and [Cartwright 2001] for further discussion of the inductive approach.

[19] I will talk freely in this paper of 'real' or 'physical' causality, signifying an objective mind-independent notion. For those who deny such a notion of cause, such talk can be recast in terms of our causal intuitions, inasmuch as there is intersubjective agreement about these. The same will apply to my talk of physical probability. See §5 on this point.

TEST evidence is obtained to confirm or disconfirm the hypothesis;

AMEND if the hypothesis is disconfirmed, the causal model is adjusted to yield a new hypothesised model;

and the process continues by returning to the prediction phase.

This approach combines aspects of both the hypothetico-deductive and the inductive methods. The inductive method is incorporated in the first stage of the causal learning process. Here AI techniques are employed to yield a causal model (or set of causal models in which case several hypotheses are evaluated simultaneously — I shall use the singular here for simplicity's sake). Thanks to the existence of the causal Markov counterexamples, we cannot be sure that the induced graph will represent the real causal relations amongst the variables. Hence the induced causal graph should be viewed as a tentative hypothesis, in need of evaluation, as occurs in the hypothetico-deductive method. Evaluation takes place in the predict and test stages. If the hypothesis is disconfirmed, rather than returning to the induce stage, local changes are made to the causal graph in the amend stage, leading to the hypothesis of a new causal theory.

Although we are primarily concerned with the general picture and its justification here, we shall take a whistle-stop tour through the stages to see how the skeleton of this method might be fleshed out.

The first stage requires a procedure for obtaining a causal model (or class of such models) from observational data, given the causal Markov condition as a basic assumption. It suffices to reiterate that inductivists have already provided several such procedures,[20] and we need neither assess these methods nor plump for any particular one here. However it is worth pointing out that the first step of the inductive procedure, namely the choice of variables that are relevant to the question at stake, is often neglected in such accounts. A good strategy here seems to be simply to observe values of as many variables in the domain of interest as possible and rule out as irrelevant those that are uncorrelated with the key variables. For example in a study to determine whether a mother's vegetarianism causes smaller babies, 105 variables related to the women's nutritional intake, health and pregnancy were measured and then the small subset of variables relevant to the key variables (vegetarianism and baby size) were determined statistically.[21]

The second step, which is also the step after the amend stage, involves drawing predictions from an induced model. Take the case in which a causal

[20] Most comprehensively to date in [Spirtes et al. 1993].

[21] [Drake et al. 1998].

model is simply a causal graph, asserting only qualitative causal relationships. One can use supposed connections between causality and probability to draw probabilistic predictions from qualitative causal claims: for example one can use the causal Markov condition to predict independence relationships, or one can apply the dictum that a cause will change the probability of its direct effect, $\mathcal{E}_{D_1,\ldots,D_k} p(e|c'D_1 \ldots D_k) - p(e|cD_1 \ldots D_k) = 0$. These predictions may not be invariable consequences of causal claims (otherwise the inductive method, and indeed a probabilistic analysis of causality, would be unproblematic) but might be expected to hold in most cases.[22] From a Bayesian perspective the confirmation one should give to causal model M given an observed failure, CMC_M say, of the causal Markov predictions from M, is proportional to $p(CMC_M|M)$, the degree to which one expects the causal Markov condition predictions to fail assuming M is correct, since Bayes' theorem gives $p(M|CMC_M) = p(CMC_M|M)\frac{p(M)}{p(CMC_M)}$. Qualitative causal claims can be used to make other plausible (but not inevitable) predictions. Causal claims are often associated with the presence of mechanisms, and one can predict that if C causes E then one will find a physical process linking C and E, with causes being spatio-temporally close to their direct effects.[23] Causality also has a close relationship with agency: manipulating causes can change effects but not vice versa.[24] Quantitative causal models lead to further predictions, either by positing local probability distributions or by asserting deterministic equational relationships.

The testing stage follows. The idea is first to collect more data — either by renewed observation or by performing experiments — in order to verify predictions made at the last stage, and second to use the new evidence and the predictions to evaluate the causal model. The hypothesised causal graph will dictate which variables must be controlled for when performing experiments. If some precise degree of confirmation is required, then, as indicated above, Bayesianism can provide this.

Finally the amend stage. It is not generally the degree of confirmation of the model as a whole which will decide how the model is to be restructured, but the results of individual tests of causal links. If, for example, the hypothesised model predicts that C causes E, and an experiment is performed which shows that changing the value of C does not change the distribution

[22]Note that a causal model resulting from the induce step will be guaranteed to satisfy the causal Markov condition, and may also be guaranteed to satisfy the probability-changing condition too, depending on the inductive algorithm. This will not be so with a causal model resulting from the amend step.

[23][Salmon 1998], [Dowe 2000].

[24][Price 1992], [Price 1992b]. See [Hausman 1998] for a collection of other asymmetries of causality, varying in plausibility, many of which may be used to generate predictions.

of E, controlling for E's other direct causes, then this evidence alone may be enough to warrant removing the arrow from C to E in the causal model. Finding out that the dependence between C and E is explained by non-causal (e.g. logical) relationships between the variables might also lead to the retraction of the arrow from C to E. The causal Markov condition may motivate adding a new common cause variable or new arrow if two variables are not found to be screened off by their current common causes. Finding physical mechanisms may suggest adding new arrows, while agency considerations may warrant changing directions of arrows. The point is that the same procedures that were used to draw predictions from a causal model may be used to suggest alterations if the predictions are not borne out.

I hope to have given enough of a glimpse of the details to see how this approach might overcome the key shortcomings of the inductive and hypothetico-deductive methods. Restructuring the hypothesised causal graph enables causal relationships to be found even in cases where the causal Markov condition fails, overcoming the key difficulty of current inductive methods. The induce and amend stages give an account of the ways in which causal theories can be hypothesised, while the predict and test stages give a coherent story as to how causal theories should be evaluated, overcoming the problem of underspecification of the hypothetico-deductive method.

§4
Justifying the Inductive Stage

There is an important justificationary question that needs addressing here: why use causal Markov methods at all in the first step of the causal learning process, given the failure of the causal Markov condition? Shouldn't the causal Markov counterexamples be reason to abandon methods that rely on this condition?

One possible response to such doubts involves arguing that although the causal Markov condition can fail in some cases, these counterexamples are rare and the condition holds most of the time. This is not a promising line of defence, however, primarily because it is very hard to execute. It is difficult to provide an estimate of the frequency of cases in which the condition holds, when the domain of application of causal reasoning is left entirely open. One may analyse cases in epidemiology and find that the condition mainly holds there, but one cannot transfer these conclusions to econometrics let alone future fields of causal research currently unknown to us.[25]

[25]Moreover there is evidence to suggest that failures of the causal Markov condition are

When faced with many counterexamples,[26] proponents of causal Markov methods tend to say that the condition would hold (i) were further 'latent' variables included in the domain of variables under consideration or (ii) were variables individuated more 'naturally'.[27] Such claims are inadequate from the point of view of a purely inductive account of causal learning because inductivists are stuck with the domain of variables that researchers initially consider relevant, individuated in ways seemingly natural at the time. It is only if we allow induced causal models to be amended post hoc that new variables can be added or their individuation changed. But this defence remains implausible until claims (i) and (ii) can be backed up by convincing arguments. Currently it is even unclear as to how to determine the referents of hypothesised latent variables, and what counts as natural individuation.

I would like to leave open the question of whether the causal Markov condition holds most of the time, as well as the question of whether the condition holds in the epistemic limit, when all relevant variables have been identified, and identified 'naturally'. Instead I will try a third type of justification of the inductive stage. I shall argue that while the causal Markov condition may fail it remains a good default assumption, in the sense that if one knows of the causal relationships amongst a set of variables, and one knows of no counterexample to the causal Markov condition amongst those variables, then one's epistemic probabilities ought to satisfy the condition. While the causal Markov condition may fail as a link between the actual causal relationships and physical probability, it should hold as a link between an agent's causal knowledge and her degrees of belief (as long as she is unaware of any failure).

The argument depends on adopting an objective Bayesian framework, and on a link between causality and probability that is somewhat less controversial than the causal Markov condition.

A Bayesian framework is required because we are concerned with the causal Markov condition as a constraint on degrees of belief. Strictly subjective Bayesianism holds that degrees of belief admit no constraints beyond the requirement that they satisfy the axioms of probability, while objective Bayesianism maintains that an agent's background knowledge should constrain her degrees of belief via the *maximum entropy principle*: subject to the constraints imposed by background knowledge, the agent's degrees of belief should be represented by the probability function that maximises the entropy $H = -\sum_v p(v) \log p(v)$, where the sum is over all assignments v of values to variables in the domain (the justification being that this function

quite widespread in practice — see [Williamson 2001], §§2,3.

[26]This does not apply to counterexamples from quantum mechanics, where arguably the causal Markov condition fails in principle, not just in practice.

[27][Spirtes et al. 1993] §3.5.1, [Pearl 2000] §2.9.1.

conveys the information encapsulated by background knowledge but as little information as possible beyond that). We shall suppose for concreteness (but without loss of generality) that our agent possesses a probabilistic causal model as background knowledge. As mentioned in §2, this contains a causal graph together with the probability distribution of each node conditional on its direct causes in the graph. These local probability distributions contain information relevant to the individual causal links, but no information bearing on the truth or falsity of the causal Markov condition.

The local probability distributions in a probabilistic causal model constrain an agent's degrees of belief in as much as her belief function should yield these distributions as marginals. It therefore remains to say how the causal graph constrains her degrees of belief. Define M_1 to be a *submodel* of M_2 if the set of variables V_1 in M_1 is a subset of the variables V_2 in M_2, the graph in M_1 is the graph induced by the graph of M_2 on V_1, and the local distributions in M_1 are those induced by the local distributions in M_2. A submodel M_1 of M_2 is *ancestral* if its domain V_1 does not include any effects of variables in $V_2 \backslash V_1$, the variables in M_2 but not in M_1, where effects are determined by the causal graph of M_2. Let p^{M_1}, p^{M_2} be the probability functions that are rational to adopt on the basis of background knowledge M_1, M_2 respectively. I propose the following principle linking causal knowledge and rational belief:

Causal Irrelevance If M_1 is an ancestral submodel of M_2 then $p^{M_2}_{\downarrow V_1} = p^{M_1}$, i.e. the restriction to V_1 of the function that is rational on the basis of M_2 is the function that is rational on the basis of M_1.

Intuitively, the causal irrelevance principle says that if an agent has initial causal knowledge M_1 and then learns of new variables that are not causes of any of the variables in M_1, thereby increasing her knowledge to M_2, then her degrees of belief about the variables in M_1 should not change. Learning of non-causes is irrelevant to current degrees of belief.

The motivation behind causal irrelevance can best be explained with the aid of an example. Suppose our agent is concerned with two variables L and B signifying lung cancer and bronchitis respectively. Initially she knows of no causal relationships between these variables. Then she learns that smoking S causes each of lung cancer and bronchitis, yielding causal graph of Figure 1. One can argue that learning of the existence of common cause S should impact on her degrees of belief concerning L and B, making them more dependent. The reasoning is as follows: if bronchitis is present, then this may be because the individual is a smoker, and smoking may also have caused lung cancer, so the agent should believe the individual has lung cancer

Figure 1: Smoking, lung cancer and bronchitis.

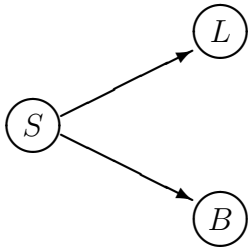
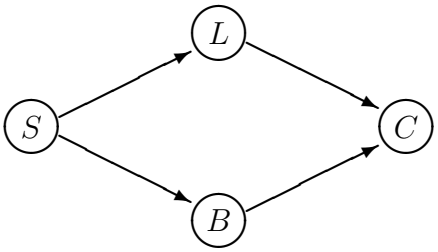


Figure 2: Smoking, lung cancer, bronchitis and chest pains.



given bronchitis to a greater extent than before — the two variables become dependent (or more dependent if dependent already).

Next the agent learns that both lung cancer and bronchitis cause chest pains C, giving Figure 2. But in this case one can *not* argue that L and B should be rendered more dependent. If an individual has bronchitis then he may well have chest pains, but this does not render lung cancer any more probable because there is already a perfectly good explanation for any chest pains. Learning of the existence of a common effect is irrelevant to the agent's current degrees of belief. Thus one cannot reason via a common effect in the same way that one can via a common cause, and the irrelevance condition picks up on this central asymmetry of causality.

The upshot of adopting the objective Bayesian framework and causal irrelevance is the following.

THEOREM 4.1 Suppose M is a probabilistic causal model and p^M is determined from M by maximising entropy subject to the constraints imposed by M, assuming the causal irrelevance principle. Then p^M satisfies the causal Markov condition with respect to the causal graph in M.

PROOF: See [Williamson 2001], §5.2. □

In fact Theorem 4.1 holds for all three types of causal model — qualitative, probabilistic and structural-equation causal models — as detailed in [Williamson 2002].

In a nutshell, then, my argument is this. Causal knowledge which includes information about individual causal relationships but no information bearing on the causal Markov condition can be represented by a causal model (qualitative, probabilistic or structural-equation). If one has such information, then one's degrees of belief ought to satisfy the causal Markov condition with respect to one's causal knowledge. Hence the causal Markov condition is justifiable when construed as an epistemic rule, linking causal knowledge with degrees of belief. If one does have knowledge that contradicts the causal Markov condition, then from the objective Bayesian point of view degrees of belief should reflect this knowledge and the causal Markov condition should not hold. Thus the causal Markov condition can hold only as a default rule, applying in the absence of counterexamples. If we accept the validity of the causal Markov condition as a default rule then we can apply it to the causal learning scenario, in which there is no initial causal knowledge, for just as the causal Markov condition can be used to constrain degrees of belief when there is causal knowledge, it can be used to constrain causal beliefs when there is probabilistic knowledge. Known causal relationships by default constrain degrees of belief via the causal Markov condition and so a hypothesised causal model should by default satisfy the causal Markov condition with respect to one's degrees of belief. These degrees of belief are informed by initial observational data, and so one can use the causal Markov condition to hypothesise a causal model from initial observational data. The default nature of the condition means that further evidence should be collected to evaluate the hypothesised causal model and the model should be amended if disconfirmed.

§5
An Analogy with Bayesianism

According to the synthesis presented in §3, a causal model is initially determined by inductive methods and is then subject to ongoing updating in the light of new evidence. There is a close parallel between this conception of causal knowledge and the objective Bayesian conception of probabilistic knowledge. According to objective Bayesianism, a belief function is initially determined via the maximum entropy principle and is then subject to ongoing updating in the light of new evidence.

It is possible to unify these two epistemic processes as follows. An agent's epistemic state contains degrees of belief together with causal beliefs. We idealise and represent these by a probability function and a directed acyclic graph respectively. Given some initial observational data, the maximum entropy principle can be used to justify the choice of prior beliefs: the maximum entropy principle determines a belief function that represents the information contained in the initial data, and also warrants the selection of a causal graph via the causal Markov condition. These prior beliefs are just beliefs — depending on the extent and reliability of initial data they may not correspond at all closely with physical probability and physical causality, in which case a process of calibration will need to take place if the beliefs are to be useful to the agent in her dealings with the world. As the agent obtains new data, mechanisms must be invoked to update prior beliefs into posterior beliefs. There are various recommendations for updating a belief function, most notably Bayesian conditionalisation and its generalisation, minimum cross entropy updating.[28] Causal beliefs are updated via some appropriate implementation of the amend stage in the scheme of §3.

This conception of belief formation and change is useful because it allows us to break a deadlock. On the one hand proponents of causal Markov learning techniques cling to a purely inductive method despite the refutation of the causal Markov condition by counterexamples, even to the point of placing the condition beyond reproach.[29] On the other hand critics of the inductive method reject causal Markov learning approaches outright on the basis of the causal Markov counterexamples. The deadlock is broken by separating the causal Markov learning techniques from the inductive method. The causal Markov counterexamples provide reason to reject the inductive method, but learning techniques that rely on the causal Markov condition remain a valuable, if fallible, way of forming causal beliefs. The Bayesian analogy pictures causal Markov learning as the first, fallible step on the path to knowledge.

The Bayesian analogy also suggests a way to defer inscrutable metaphysical questions about the nature of causality. Bayesianism has provided a purely epistemological framework in which to discuss the central issues surrounding probabilistic reasoning. By providing a degree-of-belief interpretation of probability it has been able to avoid awkward concerns about the nature of mind-independent, single-case physical probabilities and in particular how we find out about them: the epistemology of an epistemic concept of probability is not so mysterious. Likewise by providing a causal-belief in-

[28] [Williams 1980], [Paris 1994].

[29] [Pearl 2000], 44: 'this Markov assumption is more a convention than an assumption'.

terpretation of causality we do not have to answer questions about whether or how causal relationships exist as mind-independent entities, and how we can come to know about such entities. By putting the epistemology first we can deal with causality as an epistemic mind-dependent entity. We do not have to project our interpretation of nature onto nature itself, instead we can concentrate, as the prototypical inductivist Francis Bacon did, on methodology,

> our way and method (as we have often said clearly, and are happy to say again) is not to draw results from results or experiments from experiments (as the empirics do), but (as true Interpreters of Nature) from both results and experiments to draw causes and axioms, and from causes and axioms in turn to draw new results and experiments.[30]

And as with objective Bayesian probability, the epistemic view of causality does not render the concept subjective in the sense of being arbitrary or detached from worldly results:

> Human knowledge and human power come to the same thing, because ignorance of cause frustrates effect. For Nature is conquered only by obedience; and that which in thought is a cause, is like a rule in practice.[31]

Note though that the Bayesian analogy does not provide the whole story. One limitation of Bayesianism is its portrayal of the agent as a vessel receiving data, ignoring the fact that information is not just given to an agent, it must be gathered by the agent. As Popper noted, it is not enough to say to an agent 'observe' and let her get on with it — the agent must use her beliefs to narrow her search for new evidence. A picture of causal belief change must shed some light on the gathering process. Thus the predict and test stages of §3, which do not appear in the Bayesian-style conception of belief change, are of vital importance.

§6 Concluding Remarks

The syzygy of causality and probability is not as straightforward as proponents of causal Markov methods currently make out. The causal Markov

[30] [Bacon 1620] §I.CXVII.
[31] [Bacon 1620] §I.III.

counterexamples show that causal relationships need not satisfy the causal Markov condition with respect to physical probability. On the other hand, Theorem 4.1 and its generalisation show that a causal model does satisfy the causal Markov condition with respect to rational degree of belief. In [Williamson 2001] I used these two points to argue that, given a probabilistic causal model, one should beware of simply assuming the causal Markov condition and applying the resulting Bayesian network directly in probabilistic expert systems — while the Bayesian network is a good first step, one needs to calibrate the network to better represent physical probability. In this paper I use the same two points to argue a converse: given probabilistic knowledge, one should beware of simply assuming the causal Markov condition to generate a causal model — while such a model is a good first step on the way to causal knowledge, one needs to calibrate the model to better represent physical causality.

This conception of learning causal relationships yields a synthesis of the one-step inductive method and the Popperian hypothetico-deductive method. This overcomes the difficulty that the inductive method faces in learning causal relationships which do not satisfy the causal Markov condition. It also fills in some of the gaps in Popper's account of the hypothetico-deductive method. By focussing on the assignment of prior causal beliefs and their updating to posterior causal beliefs, one can give a Bayesian-style account of causal learning. We are left with a new framework for causal learning, and the somewhat daunting task of filling in the details.

References

[Bacon 1620] Francis Bacon: 'The New Organon', Lisa Jardine & Michael Silverthorne(eds.), Cambridge University Press 2000.

[Cartwright 1997] Nancy Cartwright: 'What is a causal structure?', in [McKim & Turner 1997], pages 343-357.

[Cartwright 1999] Nancy Cartwright: 'Causality: independence and determinism', in [Gammerman 1999], pages 51-63.

[Cartwright 2001] Nancy Cartwright: 'What is wrong with Bayes nets?', The Monist 84(2), pages 242-264.

[Cooper 1999] Gregory F. Cooper: 'An overview of the representation and discovery of causal relationships using Bayesian networks', in [Glymour & Cooper 1999], pages 3-62.

[Cooper 2000] Gregory F. Cooper: 'A Bayesian method for causal modeling and discovery under selection', in Proceedings of the Conference on Uncertainty in Artificial Intelligence 2000, pages 98-106.

[Corfield & Williamson 2001] David Corfield & Jon Williamson(eds.): 'Foundations of Bayesianism', Kluwer Applied Logic Series, Dordrecht: Kluwer Academic Publishers.

[Dai et al. 1997] Honghua Dai, Kevin Korb, Chris Wallace & Xindong Wu: 'A study of causal discovery with weak links and small samples', in Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan, August 23-29 1997.

[Dash & Druzdzel 1999] Denver Dash & Marek Druzdzel: 'A Fundamental Inconsistency Between Causal Discovery and Causal Reasoning', in Proceedings of the Joint Workshop on Conditional Independence Structures and the Workshop on Causal Interpretation of Graphical Models, The Fields Institute for Research in Mathematical Sciences, Toronto, Canada.

[Dowe 2000] Phil Dowe: 'Physical causation', Cambridge University Press.

[Drake et al. 1998] R. Drake, S. Reddy & J. Davies: 'Nutrient intake during pregnancy and pregnancy outcome of lacto-ovo-vegetarians, fish-eaters and non-vegetarians', Vegetarian Nutrition: an international journal 2(2), pages 45-52.

[Freedman & Humphreys 1999] David Freedman & Paul Humphreys: 'Are there algorithms that discover causal structure?', Synthese 121, pages 29-54.

[Gammerman 1999] Alex Gammerman(ed.): 'Causal models and intelligent data management', Springer.

[Glymour 1997] Clark Glymour: 'A review of recent work on the foundations of causal inference', [McKim & Turner 1997], pages 201-248.

[Glymour 2001] Clark Glymour: 'The Mind's Arrows: Bayes nets and graphical causal models in psychology', Cambridge, Massachusetts: The M.I.T. Press Press.

[Glymour & Cooper 1999] Clark Glymour & Gregory F. Cooper(eds.): 'Computation, causation, and discovery', M.I.T. Press.

[Hagmayer & Waldmann 2002] York Hagmayer & Michael R. Waldmann: 'A constraint satisfaction model of causal learning and reasoning', in Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Mahwah, NJ: Erlbaum.

[Hausman 1998] Daniel M. Hausman: 'Causal asymmteries', Cambridge University Press.

[Hausman 1999] Daniel M. Hausman: 'The mathematical theory of causation', review of [McKim & Turner 1997], British Journal for the Philosophy of Science 50, pages 151-162.

[Hausman & Woodward 1999] Daniel M. Hausman & James Woodward: 'Independence, invariance and the causal Markov condition', British Journal for the Philosophy of Science 50, pages 521-583.

[Heckerman et al. 1999] David Heckerman, Christopher Meek & Gregory Cooper: 'A Bayesian approach to causal discovery', in [Glymour & Cooper 1999], pages 141-165.

[Hempel & Oppenheim 1948] Carl G. Hempel & Paul Oppenheim: 'Studies in the logic of explanation', with Postscript in [Pitt 1988], pages 9-50.

[Humphreys 1997] Paul Humphreys: 'A critical appraisal of causal discovery algorithms', in [McKim & Turner 1997], pages 249-263.

[Humphreys & Freedman 1996] Paul Humphreys & David Freedman: 'The grand leap', British Journal for the Philosophy of Science 47, pages 113-123.

[Karimi & Hamilton 2000] Kamran Karimi & Howard J. Hamilton: 'Finding Temporal Relations: Causal Bayesian Networks versus C4.5', Proceedings of the Twelfth International Symposium on Methodologies for Intelligent System (ISMIS'2000), Charlotte, NC, USA, October 2000.

[Karimi & Hamilton 2001] Kamran Karimi & Howard J. Hamilton: 'Learning causal rules', Technical Report CS-2001-03, Department of Computer Science, University of Regina, Saskatchewan, Canada.

[Lad 1999] Frank Lad: 'Assessing the foundation for Bayesian networks: a challenge to the principles and the practice', Soft Computing 3(3), pages 174-180.

[Lemmer 1996] John F. Lemmer: 'The causal Markov condition, fact or artifact?', SIGART 7(3).

[Lewis 1986] David K. Lewis: 'Causal explanation', in [Lewis 1986b], pages 214-240.

[Lewis 1986b] David K. Lewis: 'Philosophical papers volume II', Oxford University Press.

[Mani & Cooper 1999] Subramani Mani & Gregory F. Cooper: 'A Study in Causal Discovery from Population-Based Infant Birth and Death Records', in Proceedings of the AMIA Annual Fall Symposium 1999, Hanley and Belfus Publishers, Philadelphia, PA, pages 315-319.

[Mani & Cooper 2000] Subramani Mani & Gregory F. Cooper: 'Causal Discovery from Medical Textual Data', in Proceedings of the AMIA annual fall symposium 2000, Hanley and Belfus Publishers, Philadelphia, PA, pages 542-546.

[Mani & Cooper 2001] Subramani Mani & Gregory F. Cooper: 'Simulation Study of Three Related Causal Data Mining Algorithms', in Proceedings of the International Workshop on Artificial Intelligence and Statistics 2001, Morgan Kaufmann, San Francisco, CA, pages 73-80.

[McKim & Turner 1997] Vaughn R. McKim & Stephen Turner: 'Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences', University of Notre Dame Press.

[Paris 1994] J.B. Paris: 'The uncertain reasoner's companion', Cambridge: Cambridge University Press.

[Pearl 1999] Judea Pearl: 'Graphs, structural models, and causality', in [Glymour & Cooper 1999], pages 95-138.

[Pearl 2000] Judea Pearl: 'Causality: models, reasoning, and inference', Cambridge University Press.

[Pitt 1988] Joseph C. Pitt: 'Theories of explanation', Oxford University Press.

[Pólya 1945] G. Pólya: 'How to solve it', second edition, Penguin 1990.

[Polya 1954] G. Polya: 'Induction and analogy in mathematics', volume 1 of 'Mathematics and plausible reasoning', Princeton University Press.

[Polya 1954b] G. Polya: 'Patterns of plausible inference', volume 2 of 'Mathematics and plausible reasoning', Princeton University Press.

[Popper 1934] Karl R. Popper: 'The Logic of Scientific Discovery', with new appendices of 1959, Routledge 1999.

[Price 1992] Huw Price: 'Agency and causal asymmetry', Mind 101, pages 501-520.

[Price 1992b] Huw Price: 'The direction of causation: Ramsey's ultimate contingency', Philosophy of Science Association 1992 (2), pages 253-267.

[Railton 1978] Peter Railton: 'A deductive-nomological model of probabilistic explanation', in [Pitt 1988], pages 119-135.

[Russell 1913] Bertrand Russell: 'On the notion of cause', Proceedings of the Aristotelian Society 13, pages 1-26.

[Salmon 1998] Wesley C. Salmon: 'Cauality and explanation', Oxford University Press.

[Scheines 1997] Richard Scheines: 'An introduction to causal inference', in [McKim & Turner 1997], pages 185-199.

[Spirtes et al. 1993] Peter Spirtes, Clark Glymour & Richard Scheines: 'Causation, Prediction, and Search', second edition M.I.T. Press 2000.

[Tenenbaum & Griffiths 2001] Joshua B. Tenenbaum & Thomas L. Griffiths: 'Structure learning in human causal induction', in Advances in Neural Information Processing Systems 13.

[Tong & Koller 2001] Simon Tong & Daphne Koller: 'Active Learning for Structure in Bayesian Networks', in Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence 2001, pages 863-869.

[Waldmann 2001] Michael R. Waldmann: 'Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm', Psychonomic Bulletin & Review 8, pages 600-608.

[Waldmann & Martignon 1998] Michael R. Waldmann & Laura Martignon: 'A Bayesian network model of causal learning', in M.A. Gernsbacher & S.J. Derry(eds.): 'Proceedings of the Twentieth Annual Conference of the Cognitive Science Society', Mahwah, NJ: Erlbaum, pages 1102-1107.

[Wallace & Korb 1999] Chris S. Wallace & Kevin B. Korb: 'Learning linear causal models by MML sampling', in [Gammerman 1999], pages 88-111.

[Wendelken & Shastri 2000] Carter Wendelken & Lokendra Shastri: 'Probabilistic Inference and Learning in a Connectionist Causal Network', In Proceedings of the Second International Symposium on Neural Computation, Berlin, May 2000.

[Williams 1980] Peter M. Williams: 'Bayesian conditionalisation and the principle of minimum information', British Journal for the Philosophy of Science 31, pages 131-144.

[Williamson 2001] Jon Williamson: 'Foundations for Bayesian networks', in [Corfield & Williamson 2001], pages 75-115.

[Williamson 2002] Jon Williamson: 'Maximising entropy efficiently', Electronic Transactions in Artificial Intelligence, to appear.

[Woodward 1997] James Woodward: 'Causal models, probabilities, and invariance', in [McKim & Turner 1997], pages 265-315.

[Yoo et al. 2002] Changwon Yoo, V. Thorsson & G. Cooper: 'Discovery of Causal Relationships in a Gene-regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data', in Proceedings of the Pacific Symposium on Biocomputing 2002, World Scientific, New Jersey, pages 498-509.