Synthese (2021) 199:2415–2437
https://doi.org/10.1007/s11229-020-02891-x

Conditional learning through causal models

Jonathan Vandenburgh[1]

Received: 24 April 2020 / Accepted: 23 September 2020 / Published online: 30 September 2020

Abstract
Conditional learning, where agents learn a conditional sentence 'If A, then B,' is difficult to incorporate into existing Bayesian models of learning. This is because conditional learning is not uniform: in some cases, learning a conditional requires decreasing the probability of the antecedent, while in other cases, the antecedent probability stays constant or increases. I argue that how one learns a conditional depends on the causal structure relating the antecedent and the consequent, leading to a causal model of conditional learning. This model extends traditional Bayesian learning by incorporating causal models into agents' epistemic states. On this theory, conditional learning proceeds in two steps. First, an agent learns a new causal model with the appropriate relationship between the antecedent and the consequent. Then, the agent narrows down the set of possible worlds to include only those which make the conditional proposition true. This model of learning can incorporate both standard cases of Bayesian learning and the non-uniform learning required to learn conditional information.

Keywords Conditionals · Bayesian Learning · Causal Models

1 Introduction

Suppose someone is looking for their keys in drawers A, B and C; they think each drawer is equally likely to contain the keys, so the probability that the keys are in any given drawer is $\frac{1}{3}$. Upon learning that the keys are not in drawer A, the most reasonable way to change one's beliefs is to believe that the keys are in either drawer B or drawer C, each with probability $\frac{1}{2}$. This kind of learning is successfully captured by the Bayesian theory of learning, which assumes that beliefs are represented by a

I would like to thank Fabrizio Cariani, as well as an anonymous reviewer, for helpful comments on an earlier version of this paper.

✉ Jonathan Vandenburgh
jonathanvandenburgh2021@u.northwestern.edu

[1] Northwestern University, 1880 Campus Dr., Evanston, IL 60208, USA

probability distribution and that when someone learns a proposition A, their beliefs change from some prior distribution Pr to a posterior distribution $\Pr_A$ according to Bayes' Theorem.[1] The posterior distribution $\Pr_A$ is given by conditionalization, so for any proposition B, $\Pr_A(B) = \Pr(B|A) = \frac{\Pr(B \wedge A)}{\Pr(A)}$. Bayesian learning therefore predicts that learning is uniform, so the same updating procedure applies for any proposition in any situation.

Now suppose the information one learns is in the form of a conditional sentence 'If A, then B.' While one might expect conditional learning to be uniform in the same way Bayesian learning is, many examples (Douven 2012) suggest this is not the case. Consider, for example, the conditionals 'If my brother is here, the keys are in drawer C' and 'If the keys are in drawer A, then someone moved them.' In the first example, the conditional should not change the credence that the speaker's brother is here, but may affect the probability that the keys are in drawer C. In the second example, the conditional seems to express the speaker's belief that the keys were not originally in drawer A; if it is unlikely someone moved them, then the probability that the keys are in A should decrease. Thus, in one case, the conditional does not affect the probability of the antecedent, but in the other case, it decreases the probability of the antecedent.

Many theories of conditional learning fail to predict this lack of uniformity. The most popular approach to conditional learning has been to assume that people learn conditionals through Bayesian updating on the material conditional $A \supset B$, or $\neg A \vee B$. This has nice theoretical properties; this approach to conditional learning produces a posterior distribution which minimizes the Kullback-Leibler divergence between the prior and the posterior (Van Fraassen 1981). However, this procedure always decreases the probability that the antecedent is true, offering counterintuitive predictions in some cases (Douven and Romeijn 2011; Douven 2012). For example, this approach cannot explain the intuition above that some conditionals leave the antecedent probability unchanged. Nevertheless, this view retains some advocates; Eva et al. (2020), for example, argue that they can use material conditional learning to account for the observed variations in the posterior by considering other propositions learned alongside the conditional propositions. Another proposal (Bradley 2005) predicts that the antecedent probability always remains constant, failing to explain cases where the antecedent probability changes. Other authors, such as Huisman (2017), take the observed lack of uniformity in conditional learning as evidence that conditional learning cannot be incorporated within the Bayesian framework.

This paper proposes a causal model for conditional learning which can explain the observed variations in the posterior distributions. Sect. 2 provides a formal account of Bayesian learning based on a possible worlds semantics for propositions. Section 3 introduces a causal constraint governing conditional learning, arguing that the posterior probability of the antecedent depends on the causal structure behind the conditional. This goes beyond the constraints introduced by Douven (2012), offering a systematic explanation for different posterior judgments in different contexts. Furthermore, the conditionals introduced here do not rely on additional information in the background.

Section 4 introduces causal models and presents a generalization of Bayesian learning to causal belief states, showing that this coincides with ordinary Bayesian learning

[1] For psychological evidence in favor of Bayesian reasoning, see Oaksford and Chater (2007).

for non-conditional propositions. Section 5 introduces a causal semantics for conditionals and shows how the problem of conditional learning arises again in the semantics of left-nested conditionals. Section 6 presents a causal approach to conditional learning. On this theory, conditional learning proceeds in two steps: first, one learns a new causal model, if necessary, which ensures that the antecedent and the consequent are probabilistically dependent, and second, one performs Bayesian updating on the largest set of possible worlds which makes the conditional true in the new causal model. This, I show, can explain the predictions identified in Sect. 3. Section 7 presents some further examples of conditional learning within the causal framework, including the 'common cause' conditional and examples from Douven (2012) where background information becomes relevant.

The main competing account which rivals the predictions of this paper is an approach which uses Jeffrey imaging on Stalnaker conditionals (Günther 2017, 2018). Since this approach builds on the Stalnaker semantics for conditionals, the predictions rely on judgments about which worlds are most similar to which other worlds. While this can account for the causal constraints presented here, it can only do so if the similarity ordering respects the background causal structure, lending explanatory priority to the causal account. Additionally, the account presented here is consistent with any conditional semantics which respects causal constraints on the background set of worlds, providing a more general theory that does not rely on Stalnaker's more restrictive conditional semantics. Furthermore, the causal constraints on conditional learning in Sect. 3 and the causal generalization of Bayesian learning in Sect. 4 make contributions independent of the model of conditional learning in Sect. 6.

2 Bayesian learning

In Bayesian models of learning, agents' epistemic states are represented by credence functions on propositions. Learning a proposition A leads the agent to update his or her credence function by conditionalization. Thus, if $\mathcal{A}$ is a field of propositions, we assume that the agent has a credence function $P : \mathcal{A} \to [0, 1]$ such that (i) for any tautology T, $P(T) = 1$ and (ii) for mutually exclusive A and B, $P(A \vee B) = P(A) + P(B)$. When an agent learns a proposition $A \in \mathcal{A}$, the agent forms a new credence function P_A which assigns probability 1 to proposition A. This credence function is given by conditionalization, so for any belief B, $P_A(B) = P(B|A) = \frac{P(B \wedge A)}{P(A)}$.[2]

Identifying propositions with sets of possible worlds will facilitate the discussion of conditional learning and learning in causal models. In this case, we model belief states as probability distributions over a set of possible worlds. We assume that we have a set of possible worlds Ω and that for each proposition $A \in \mathcal{A}$, A represents a set of possible worlds, $[A] \subseteq \Omega$. We assume that Ω is a probability space, so there is a sigma algebra $\Sigma \subseteq \mathcal{P}(\Omega)$ of measurable subsets of Ω and a probability measure Pr on Σ. To simplify exposition, I will assume Ω is finite throughout, though the discussion extends to the infinite case. When Ω is finite, all subsets of Ω are in Σ

[2] For an introduction to Bayesian epistemology, see Hartmann and Sprenger (2010).

and a probability distribution Pr is fully determined by an assignment $\Pr(\omega)$ to each $\omega \in \Omega$ such that:

(i) for all $\omega \in \Omega$, $\Pr(\omega) \in [0, 1]$
(ii) $\sum_{\omega \in \Omega} \Pr(\omega) = 1$.

Any proposition A is associated with a set of possible worlds $[A] \subseteq \Omega$, which we assume is measurable, so $[A] \in \Sigma$. The likelihood that A is true, $\Pr(A)$, is determined by the likelihood that the world is an A-world. For finite Ω, this means that

$$\Pr(A) = \sum_{\omega \in [A]} \Pr(\omega).$$

It is not hard to see that this credence assignment satisfies the conditions for a credence function stipulated above.

The possible worlds formalism also allows us to handle belief updating by conditionalization. Learning the proposition A corresponds to restricting the space of possible worlds Ω to the set of A-worlds, $[A]$, and updating the probability distribution accordingly: for any $B \in \mathcal{A}$ (or, more generally, $B \in \Sigma$), the new distribution $\Pr_A$ is given by $\Pr_A(B) = \Pr(B|A) = \frac{\Pr(B \wedge A)}{\Pr(A)}$. For a world ω, the updated probability $\Pr_A(\omega)$ is 0 in worlds where A is false and $\frac{\Pr(\omega)}{\Pr(A)}$ in worlds where A is true. Since the new probability function $\Pr_A$ yields a probability distribution over possible worlds, we can think of the new belief space as the set of possible worlds $\Omega_A = [A]$ with probability distribution $\Pr_A$. Thus, Bayesian learning predicts that, given a belief space consisting of a set of possible worlds Ω with probability distribution Pr, learning A results in a new, restricted, belief space Ω_A with a new probability distribution $\Pr_A$ given by conditionalization. The characterization of Bayesian learning in terms of possible worlds will prove useful in understanding how to extend the theory to incorporate conditional sentences.

3 Constraints on conditional learning

A problem for Bayesian learning is that there is no straightforward extension to the case of conditional learning, where one learns a conditional 'If A, then B'. The above account of Bayesian learning proposes a method for learning propositions, or sets of possible worlds. However, there is no straightforward proposition, or set of possible worlds, corresponding to the conditional. Adams (1975) and Edgington (1995), for example, argue that the conditional does not have truth conditions, Kaufmann (2001), Bradley (2002) and Rothschild (2014) argue that the truth conditions of conditionals are not binary, and the dominant linguistic approach to conditionals, following Kratzer (1986, 2012), treats the conditional as a restricted modal rather than a stand-alone proposition.

Despite the difficulties in settling on a propositional account of the conditional, some authors (Van Fraassen 1981; Eva et al. 2020) have proposed that people update their beliefs according to the material conditional, $A \supset B$, even though the material conditional does not work in general as a semantic theory of the conditional. This

approach has some theoretical virtues: for example, it minimizes the Kullback–Leibler divergence between the prior and the posterior distribution, generalizing in one way Bayesian updating. However, this approach leads to counterintuitive predictions for how people learn conditional information.

The main issue is that, since $A \supset B = \neg A \vee B$, learning $A \supset B$ only eliminates worlds where A is true, so learning a conditional always decreases the probability that the antecedent is true. However, there are many examples where this seems absurd. Consider the sundowners example from Douven and Romeijn (2011), where one learns the conditional 'If it rains tomorrow, sundowners (a party) will be canceled.' Suppose we initially think rain and sundowners are independent and both have 50% chance of occurring. Then, if we let R stand for rain and S stand for sundowners, our epistemic state consists of four possible worlds, $R \wedge S$, $R \wedge \neg S$, $\neg R \wedge S$, $\neg R \wedge \neg S$, each with equal probability $\frac{1}{4}$ of occurring. Learning $R \supset \neg S$ eliminates $R \wedge S$ and leads to Bayesian updating on the remaining worlds, which now have equal probability of $\frac{1}{3}$. This means that $\Pr(R) = \Pr(R \wedge \neg S) = \frac{1}{3}$, so the probability of rain decreases from $\frac{1}{2}$ to $\frac{1}{3}$. This is problematic because we think of the conditional as telling us something about the relationship between rain and sundowners, or what will happen when it rains, but not communicating any information about whether it will rain or not.

In light of this concern, some authors have proposed slight modifications of this approach to conditional learning. Bradley (2005), for example, has argued that we learn a conditional 'If A, then B' by updating on both the material conditional $A \supset B$ and the constraint that the probability of the antecedent remains fixed. Thus, in the sundowners example, the new credences are $\Pr(R \wedge \neg S) = 0.5$ and $\Pr(\neg R \wedge S) = \Pr(\neg R \wedge \neg S) = 0.25$, a much more reasonable prediction than what follows from just updating on the material conditional. However, this is inconsistent with cases highlighted by Douven (2012) which suggest that learning a conditional can lead the antecedent probability to decrease, increase, and remain the same. While Douven's cases rely on background explanatory concerns and will be discussed in greater detail in Sect. 7, we can illustrate the point by considering another example where the probability of the antecedent decreases.

Suppose one is receiving a test (T) for a disease (D), where one initially thinks the likelihood of having the disease is 10% and the test comes back positive in all cases where one has the disease, as well as 10% of the time when one is healthy (a false positive). Consider learning the conditional 'If the test comes back positive, the disease is present.' It seems that we are learning that the test has no false positives, so $\Pr(T \wedge \neg D) = 0$. Thus, we expect the probability of having the disease to stay the same (since we did not learn anything about its presence) and the probability that the test is positive to decrease from 20% to 10%. This coincides with the prediction of material conditional learning, while Bradley's account predicts that $\Pr(T)$ stays the same at 0.2 and $\Pr(D)$ increases to 0.2, making the unfortunate prediction that a more accurate test leads to a higher likelihood of disease.

Thus, in the sundowners case, we expect the antecedent probability to stay the same, while in the disease testing case, we expect the antecedent probability to decrease. Possible explanations for this include background effects on the explanatory status of the antecedent (Douven 2012), background information learned in addition to the

conditional (Eva et al. 2020), and which antecedent world is closest to the actual world (Günther 2018). The most fundamental difference, I argue, comes from the difference in causal structure underlying the two examples.

In the sundowners example, rain has a causal effect on the cancelation of sundowners, so the conditional is causal: it communicates that the antecedent has a causal effect on the consequent. In the disease testing example, the positive test is a causal result of the disease, so the conditional is diagnostic: the antecedent is a causal result of the consequent.[3] Intuitively, we expect a difference in how one learns causal and diagnostic conditionals. For a causal conditional 'If A, then B,' where A causes B, we do not learn anything about A, but learn that B must always be the case conditional on A, so we expect the probability of A to remain the same and the probability of B to increase. For a diagnostic conditional 'If A, then B,' where B causes A, we learn that B is the only cause of A, so the likelihood of B stays the same while the likelihood of A decreases since we have ruled out causes of A other than B.

We can explain the difference in learning causal and diagnostic conditionals by formalizing the learning process within causal models. While the formal procedure will be presented in Sect. 6 after a discussion of causal models and conditional truth conditions, at this point we can outline the intuition for how causal models can explain these cases. Conditional learning through causal models proceeds in two steps: (1) learning a causal model which can explain the relationship between the antecedent and the consequent and (2) ruling out states of the world which are inconsistent with the conditional information.

In the sundowners example, the causal information we learn is that rain causes the cancelation of sundowners. We can formalize learning this causal information when we put the example into the terminology that will be introduced formally in Sect. 4. We begin with two endogenous variables of interest, rain (R) and sundowners (S), which are governed by exogenous variables U_R and U_S, so $R = U_R$ and $S = U_S$, where U_R and U_S are independent. The exogenous variables U_R and U_S represent whatever background factors cause rain and sundowners, which are not theorized about further in the model. When we learn the conditional, we learn that this initial causal model is insufficient to represent the relationship between R and S since R and S are independent in the model. Thus, to incorporate the new information, we must form a new causal structure of the world where R has a causal effect on S, which we can represent as a graph on endogenous variables:

$$R \downarrow S.$$

This graph comes with a new set of 'structural equations' which determine the relationships between variables. We keep R exogenous, so $R = U_R$, but now sundowners is canceled in the case of rain, so $S = \neg R \wedge U_S$: sundowners needs an absence of rain to occur. If we keep our prior constant on the exogenous variables

[3] The division of conditionals into causal and diagnostic conditionals also plays a role in van Rooij and Schulz (2019).

in the model, so $\Pr(U_R) = \Pr(U_S) = 0.5$, then in the new model, $\Pr(R) = 0.5$ and $\Pr(S) = \Pr(S \wedge R) + \Pr(S \wedge \neg R) = 0 + 0.25 = 0.25$. The fact that the probabilities of U_R and U_S stay the same represents the fact that we haven't learned anything about rain or the causes underlying sundowners absent rain; we have only learned about what happens to sundowners in the case of rain. This causal updating procedure, where we change the relationships between variables but keep all other beliefs about these variables constant, predicts that the probability of rain remains constant, agreeing with the posterior which results from Bradley's account. This analysis also extends beyond the sundowners case to any causal conditional 'If A, then B', where we can interpret the conditional as 'A causes B' or 'A leads to B'. In these cases, the probability of the antecedent remains constant and the probability of the consequent increases.

In the sundowners case, creating the new causal model for the situation is sufficient to learn the conditional information: the structural equation $S = \neg R \wedge U_S$ rules out the possibility of $R \wedge S$, so no situations prevent $\neg S$ from being true whenever R is true. This means step (1) is sufficient to learn the conditional information with no need to rule out further possibilities from step (2). The disease testing case, on the other hand, illustrates a case where we already have the correct causal model, but must rule out some of the possible worlds.

Recall that we begin in the disease testing case by understanding that the disease (D) causes a positive test result (T), where $\Pr(D) = 0.1$ and a positive test result occurs whenever the disease is present and 10% of the time when the disease is absent, so $\Pr(T) = 0.2$. Here, the causal structure is as follows, both before and after learning the conditional:

$$D \downarrow T.$$

Prior to learning the conditional, the structural equations governing the causal relationships are $D = U_D$ and $T = D \vee U_T$. Here, U_T stands for any cause of T other than D, which represents the false positives of the test T. When learning the conditional, we learn that D is the only cause of T, so we should eliminate the possibility that the test will come out positive for a reason other than the disease D, which corresponds to U_T. This leads to updated structural equations $D = U_D$ and $T = D$, where the possibility U_T is no longer relevant. Since nothing changed about D, the probability of D remains the same ($\Pr(D) = 0.1$), but T occurs in fewer circumstances than before, so now $\Pr(T) = 0.1$ rather than 0.2; we learn, as expected, that the test does not have false positives, but our beliefs about the presence of the disease remain the same.

The two examples presented here highlight a problem for accounts of conditional learning. How we update a probability distribution, and specifically how we determine the updated probability of the antecedent, changes in different circumstances. Furthermore, whether the antecedent is causally relevant for or causally dependent on the consequent seems relevant for how we update our beliefs based on the conditional. For a causal conditional, where the conditional tells us how the consequent changes given the antecedent, the antecedent probability remains constant. For a diagnostic con-

ditional, where the conditional tells us in which circumstances the antecedent arises from the consequent, the antecedent probability decreases. These examples also offer a preview for how a causal approach can account for these cases: the structural equations governing the causal relations determine which changes are required to make the conditional true, explaining why different posterior distributions result in different cases. This motivates the introduction of a causal approach to learning in Sect. 4 and its extension to conditionals in Sect. 6.

4 Causal structures and Bayesian learning revisited

To extend Bayesian learning to handle causal information, we will introduce the notion of a causal model. Within a causal model, the exogenous variables determine all of the information represented in the model. A causal model paired with an exogenous variable assignment, then, will determine a causal world; the causal worlds act as truthmakers for propositions expressed within the causal model. We can then use this notion of a causal world to extend the approach to epistemic states and learning presented in Sect. 2: an epistemic state is a probability distribution over causal worlds and learning a proposition corresponds to limiting the set of causal worlds and updating the probability distribution accordingly.

To make this discussion of learning more precise, we start by defining a causal model.[4] A causal model $\mathcal{M} = (U, V, f_i)$ requires a finite set of exogenous variables U, a set of endogenous variables $V = (V_1, \ldots, V_n)$, and a set of structural equations $F = (f_1, \ldots, f_n)$ determining the values of the endogenous variables. For each i, the structural equation f_i determines the value for V_i as $v_i = f_i(pa_i, u_i)$, where pa_i is a value for the parents PA_i of V_i and u_i is an assignment to the exogenous variables needed to determine V_i. The set of parents PA_i for each endogenous variable V_i determines a directed acyclic graph (DAG) $\mathcal{G}$ over V, where we draw an arrow from V_i to V_j if V_i is a parent of V_j. Since all structural equations depend on a variable's parents and exogenous variables, and the endogenous variables without parents are determined completely exogenously, setting the exogenous variables completely fixes the endogenous variables of the model. This means that the structural equations determine the values of V given values of U, so the set of structural equations forms a function $F : U \rightarrow V$.

As discussed above, a causal world $(\mathcal{M}, u)$ is a causal model $\mathcal{M}$ paired with an exogenous variable assignment $u \in U$. The propositions of interest in causal models are logical combinations of variable assignments, generally restricted to the endogenous variables. For a variable assignment $V_i = v_i$, we get a set of possible worlds $[V_i = v_i] = \{u \in U : F(u)_i = v_i\} \subseteq U$, where $V_i = v_i$ is true in a world u when plugging u into the structural equations F makes V_i take on value v_i. Since individual variable assignments yield sets of possible worlds, we can associate logical combinations of variable assignments with sets of possible worlds through the Boolean

[4] This presentation of causal models follows Vandenburgh (2020). For a more detailed discussion of causal models, see Pearl (2009).

operations: conjunction corresponds to set intersection, disjunction to set union, and negation to complementation.

To see an example illustrated more formally, recall the above case where a test T tests for disease D with false positives. Here, we have two exogenous variables, U_D determining the presence of the disease and U_T representing any cause of a positive test result other than D, and D and T are our endogenous variables. The structural equations in this model are $D = U_D$ and $T = D \vee U_T$. Since the variables are binary, this model has four possible worlds, corresponding to the four states of the exogenous variables U_D and U_T. We can consider propositions built from endogenous variable assignments, such as 'The disease is present or the test comes back positive,' $D \vee T$. This proposition is true in three of the four possible worlds, only false when $U_D = 0$ and $U_T = 0$.

Now that we have defined causal worlds and propositions, we can define an epistemic state as a probability distribution Pr over the set of causal worlds. Taking the causal structure as given, this is simply a probability distribution over exogenous variable assignments. We assume that the exogenous variables are probabilistically independent of each other, so when $i \neq j$, $\Pr(U_i = u_i | U_j = u_j) = \Pr(U_i = u_i)$. Causal models with probability distributions satisfying this condition are called Markovian models. Note that a distribution on U leads to a distribution over V: for $v \in V$, $\Pr(v) = \Pr(\{u \in U : F(u) = v\})$; we will denote $\{u \in U : F(u) = v\}$ as U_v. If A is a proposition, so $[A] \subseteq U$, we can use the Bayesian learning mechanism from Sect. 2 to learn A within the causal model: we restrict to the set U_A of worlds where A is true and use the Bayesian updated probability distribution $\Pr_A$.

We can see that this learning procedure always agrees with the result we get if we ignore the causal model and use Bayesian updating on the endogenous variables. Suppose we have a distribution Pr on V and a distribution π on U so that Pr agrees with the distribution π induces on V : $\Pr(v) = \pi(U_v)$. To see that traditional Bayesian updating and Bayesian updating within the causal model give the same results, we must verify that $\pi_A(U_v) = \Pr_A(v)$ for any $A \in \mathcal{A}$ and for any $v \in V$. To demonstrate this, first note that the proposition A defines both a set of worlds $[A]_U \subseteq U$ and a set of worlds $[A]_V \subseteq V$. For all u such that $F(u) = v$, the same propositions are true of v and u, so $v \in [A]_V$ iff, for all u such that $F(u) = v$, $u \in [A]_U$. Then we can see that A has the same probability according to both π and Pr:

$$\pi(A) = \sum_{u \in [A]_U} \pi(u) = \sum_{v \in [A]_V} \pi(U_v) = \sum_{v \in [A]_V} \Pr(v) = \Pr(A).$$

Then for any $v \in V$, when A is not true in V, $\Pr_A(v) = \pi_A(v) = 0$, and when A is true in V, $\Pr_A(v) = \frac{\Pr(v)}{\Pr(A)} = \frac{\pi(U_v)}{\pi(A)} = \pi_A(U_v)$. This shows that Bayesian updating within a causal model produces the same result that Bayesian updating would produce if we ignored the causal model.

We can illustrate this with an example. Consider the testing example with false positives. In the possible worlds framework from Sect. 2 without a causal model, we have three possible worlds with probability distribution: $\Pr(D \wedge T) = 0.1$, $\Pr(\neg D \wedge T) = 0.1$, and $\Pr(\neg D \wedge \neg T) = 0.8$. To translate this into the causal framework above, we need a probability distribution π on exogenous variables which induces Pr. It is not

hard to see that this is satisfied by the distribution where $\pi(U_D) = \frac{1}{10}$ and $\pi(U_T) = \frac{1}{9}$, where U_D and U_T are independent, satisfying the Markov condition. We can calculate the full distribution using independence, so $\pi(U_D \wedge U_T) = \frac{1}{90}$, $\pi(U_D \wedge \neg U_T) = \frac{4}{45}$, $\pi(\neg U_D \wedge U_T) = \frac{1}{10}$, and $\pi(\neg U_D \wedge \neg U_T) = \frac{4}{5}$. We can see that π induces Pr on V.

As an example, consider the proposition T, that the test is positive. In the original representation, $\Pr(T) = \Pr(D \wedge T) + \Pr(\neg D \wedge T) = 0.2$. In the causal representation, a positive test result comes from D whenever the disease is present and from U_T when the disease is not present since $T = D \vee U_T$, so $\pi(T) = \pi(U_D) + \pi(\neg U_D \wedge U_T) = 0.1 + 0.1 = 0.2$. We can see that updating by T leads to the same distribution for π in the causal model and for Pr without the causal model. For Pr, we get that there are two worlds, $\Pr_T(D \wedge T) = 0.5$ and $\Pr_T(\neg D \wedge T) = 0.5$. For π, we have three worlds: $\pi_T(U_D \wedge U_T) = \frac{1}{18}$, $\pi_T(U_D \wedge \neg U_T) = \frac{4}{9}$, $\pi_T(\neg U_D \wedge U_T) = \frac{1}{2}$. We can see that π_T and $\Pr_T$ are equivalent on endogenous variables: $\pi_T(D) = \Pr_T(D) = 0.5$, for example. This shows that Bayesian learning of non-conditional propositions can be incorporated into the causal modeling framework; the main advantage of causal models will come in the next sections when we discuss conditional truth conditions and how to learn conditional information.

5 Conditional truth conditions and the problem of left-nested conditionals

In order to find a procedure for updating beliefs to make a conditional true, we must discuss the truth conditions for conditionals. As discussed in Sect. 3, this is a complex issue with little agreement. Here, I will briefly argue for a causal interventionist theory of conditional truth conditions, defended in greater depth in Vandenburgh (2020). This theory reconciles two competing approaches to the truth conditions of counterfactuals: the possible worlds approach (Lewis 2013; Stalnaker 1968; Kratzer 1986) and the causal modeling approach (Pearl 2009; Hiddleston 2005).

The causal modeling approach defines conditional truth conditions relative to a causal model using causal concepts like interventions and minimal changes. Using causal models allows us to avoid less formal notions like similarity or relevance and makes it easier to incorporate causal intuitions relevant for conditional judgment, like those presented in Sect. 3. However, popular theories which rely on causal models face two disadvantages relative to possible worlds theories: they fail to generalize to logically complex conditionals (for example, conditionals with disjunctive antecedents) and they do not clearly correspond to well-known conditional logics.[5] The possible worlds approach takes a conditional 'If A, then B' to be true if B is true in all relevant worlds where A is true, where the set of relevant A-worlds depends on the actual world u and is represented by the selection function $f(A, u)$. The use of possible worlds makes the extension to logically complex conditionals straightforward and allows us to easily verify axioms of conditional logic (Nute and Cross 2001; Arlo-Costa 2019). However, the selection function is usually determined by informal notions like simi-

[5] The original theories of Hiddleston and Pearl only apply to a limited set of counterfactual antecedents. Briggs (2012) offers one extension of Pearl's model to more complex antecedents, but this comes with undesirable logical consequences: modus ponens, for example, is not valid.

larity and relevance rather than more robust causal notions. The theory presented here combines these two approaches: it defines conditional semantics in terms of a selection function, which has nice logical properties, but uses causal criteria to determine the worlds in the selection function. On this theory, the selection function $f(A, u)$ is the set of worlds where we perform some causal intervention on the actual world u to set A true. Here, I implement a strict semantics which incorporates all causally relevant interventions and satisfies the axioms of Pollock's (1981) logic **SS**. However, one could easily restrict the selection function further based on additional notions of similarity, relevance, or minimal change, and therefore get a stronger logic such as Lewis's **VC** or Stalnaker's **C2**, without changing the causal framework or significantly impacting the results for conditional learning.

It is worth noting that, while causal models are typically used for counterfactual rather than conditional semantics, there is good reason to believe that the model carries over to indicative conditionals. In many cases, indicative and counterfactual conditionals have the same meaning: consider, for example, 'If it rains, sundowners will be canceled' and 'If it were to rain, sundowners would be canceled.' Furthermore, many semantic theories, such as those of Lewis and Stalnaker, are applied to both indicative and counterfactual conditionals.[6] However, there are also well known cases where judgments about indicative and counterfactual conditionals come apart: contrast 'If Oswald didn't shoot Kennedy, someone else did' with 'If Oswald hadn't shot Kennedy, someone else would have' (Adams 1970). While understanding the differences between indicative and counterfactual conditionals poses an interesting research question relevant for the causal modeling of conditional truth conditions and conditional learning, this falls beyond the scope of the present study.[7]

To define the truth conditions of a conditional 'If A, then B' in a world u, we must define the selection function $f(A, u)$, or the set of causally relevant A-worlds. To do this, we consider all of the minimal 'interventions' one could make in the world to set A true.[8] An intervention is a change to some of the exogenous variables which produces a causal effect but leaves all aspects of the world independent of the change constant.[9] Restricting ourselves to the minimal interventions eliminates the need to consider worlds where we change variables not necessary to set the antecedent true; for example, if we are interested in changes to the weather, we should not consider what happens if we change someone's shopping preferences. The set of relevant A-worlds is then the set of all worlds where one has applied a minimal intervention to set A true in u.

Suppose the causal model has m exogenous variables, $U = (U_1, \ldots, U_m)$. A restricted variable assignment r is an assignment to some subset of these variables, $S \subseteq \{1, \ldots, m\}$. If U_S represents the assignments to variables in U corresponding to indices S, then an assignment r to variables S is an element $r \in U_S$. Given a world u,

[6] Note, however, that Lewis himself opposed applying his semantics for counterfactuals to indicatives; see Lewis (2013).

[7] See also Weatherson (2001) and Khoo (2015).

[8] There is some experimental support for using minimal interventions; see, e.g., Rips (2010).

[9] Note that typical definitions of an intervention allow for interventions on endogenous variables to make the variables independent of their parents (Pearl 2009; Hagmayer et al. 2007; Fisher 2017a). In the theory used here, we restrict interventions to the exogenous variables.

we can intervene with the restricted assignment r by setting the values of u to r on S and leaving u unchanged outside of S : $u|r = r \times u|_{S'}$, where S' is the complement of S.

This allows us to define the set of all restricted variable assignments which set the antecedent A true for a given world u:

$$R_u(A) = \{r : \exists S, r \in U_S \ \& \ u|r \in [A]\}.$$

This is just the set of restricted assignments r where A is true in $u|r$. This set is non-empty as long as A is true in some world: if $w \in [A]$, then $w \in R_u(A)$ for all u since $u|w = w$.

As discussed above, we do not want to include all variable assignments which set A true, since these involve changing variables which are irrelevant to A. Instead, we want to restrict the set $R_u(A)$ to include only those variable changes which are necessary to set A true. We notice that if i is such a minimal intervention, then any extension of i to other variables would keep A true, so any extension would also be an element of $R_u(A)$. We can formalize this minimality condition by defining an order $\leq$ on $R_u(A)$: if $r_1, r_2 \in R_u(A)$ make assignments to S_1 and S_2, respectively, then $r_1 \leq r_2$ iff $S_1 \subseteq S_2$ and $r_2|_{S_1} = r_1$. This is satisfied when r_2 is an extension of r_1: r_2 makes the same assignment on all variables r_1 changes, but may change more variables.

This allows us to define the set of minimal interventions which force A, $I_u(A)$: $i \in I_u(A)$ if i is an $\leq$-minimal element of $R_u(A)$. Formally,

$$I_u(A) = \{i \in R_u(A) : \nexists r \in R_u(A), r \neq i, r \leq i\}.$$

We can use this set of interventions to define the selection function for conditional semantics: $f(A, u) = \{u|i : i \in I_u(A)\}$. This selection function allows us to define truth conditions for the conditional: a conditional 'If A, then B,' written $A \rightarrow B$, is true in a world u if B is true in all intervened worlds in $I_u(A)$. Thus, whenever we make a minimal intervention to set A true, B must be true. This gives a set of worlds where the conditional $A \rightarrow B$ is true:

$$[A \rightarrow B] = \{u \in U : \forall i \in I_u(A), u|i \in [B]\}.$$

To see how the truth conditions for conditionals work, consider again the example where a test T tests for disease D with false positives. Recall that there are two exogenous variables, U_D determining the presence of the disease and U_T representing any cause of a positive test result other than D, with structural equations $D = U_D$ and $T = D \vee U_T$. Consider the conditional $D \rightarrow T$: 'If the disease is present, the test is positive.' Here, any variable assignment which sets $U_D = 1$ sets the antecedent true (formally, is in $R_u(A)$ for any u) and the assignment setting $U_D = 1$ is the unique minimal element of $R_u(A)$ for every world u (the only element of $I_u(A)$). For every u, setting $U_D = 1$ guarantees that $T = 1$ since $D = U_D$ and $T = D \vee U_T$, so the conditional $D \rightarrow T$ is true in every world. Now consider the conditional $T \rightarrow D$: 'If the test is positive, then the disease is present.' This is the diagnostic conditional considered in Sect. 3. The two possible minimal interventions which set

the antecedent true are $U_D = 1$ and $U_T = 1$; the only worlds where the minimal interventions guarantee that $D = 1$ are those where $U_D = 1$ already, or where the consequent is already true. When the consequent is not true, the possibility of false positives makes the diagnostic conditional false.

Having defined truth conditions for conditionals, we can identify another problem which arises for conditional learning. Since we defined a set of possible worlds associated with the conditional, $[A \rightarrow B]$, we might expect to get good predictions for the truth conditions of left-nested conditionals. However, this is not the case. Consider the conditional 'If the disease is present whenever the test is positive, then the disease is present,' $(T \rightarrow D) \rightarrow D$. As we saw above, the conditional $T \rightarrow D$ is true only in worlds where $U_D = 1$, so this is the minimal intervention we need to consider for the antecedent $T \rightarrow D$. However, the intervention $U_D = 1$ always guarantees that $D = 1$; therefore, this left-nested conditional is predicted to be true in every world. However, this is absurd: learning that the test is only positive when the disease is present should eliminate the chance of false positives, not teach us that the consequent is true (that the disease is present). This is a problem which arises for left-nested diagnostic conditionals: the closest worlds where the conditional is true are predicted to be the worlds where the consequent is true, a very counterintuitive result.

This shows that treating a left-embedded conditional as a set of possible worlds, on the semantic theory presented here, yields counterintuitive truth conditions. Part of the reason for this is that the conditional proposition may require changes to the causal model itself, as was the case in the sundowners example in Sect. 3. Another reason, as illustrated here, is that conditional truth conditions depend on both the actual world and the selection function, or the possible worlds that are relevant for evaluating the conditional. Sometimes, conditional learning requires not just that we impose a constraint on the world to make the conditional true, but that we impose a constraint on the set of possible worlds that will enter into the selection function. This is the case in the above example, where the conditional seems to communicate that false positives are no longer possible, so we should remove $U_T = 1$ from the set of relevant possibilities.

This motivates the account of conditional learning introduced in the next section, which will be able to handle learning both causal and diagnostic conditionals. This learning will require two steps: first, we must find a causal model which connects the antecedent and the consequent, and second, we must narrow down the space of possibilities in a way which makes the conditional true in every possible world.

6 Conditional learning in causal models

We now return to the problem of conditional learning. Recall from Sect. 2 how Bayesian learning of a proposition A works: we start with a belief state, which is a probability distribution Pr over a set of possible worlds Ω, and after learning A, we have an updated belief state consisting of a smaller set of possible worlds Ω_A and a new probability distribution $\Pr_A$ over Ω_A defined through Bayesian updating. Section 3 argued that conditional learning depends on the causal structure underlying the conditional, Sect. 4 introduced a causal model for an agent's epistemic state using a probability distribution over causal worlds, and Sect. 5 defined truth conditions

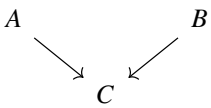
for conditionals in a causal world. In defining a procedure for conditional learning, we want a procedure which takes as input (1) a belief state consisting of a causal model $\mathcal{M} = (U, V, f_i)$ and a probability distribution Pr over U and (2) a conditional $A \rightarrow B$, where A and B are propositions formed from variable assignments in V, and outputs a new causal model $\mathcal{M}' = (U', V, f_i')$, where $U' \subseteq U$, and a new probability distribution Pr$'$ over U' such that $A \rightarrow B$ is true in all worlds in $\mathcal{M}'$.

Two changes are necessary for the new causal model $\mathcal{M}'$ to ensure that the conditional $A \rightarrow B$ is true: first, the causal graph and structural equations must allow for the right dependence between B and A and, second, we must eliminate any possible worlds from U which prevent A from guaranteeing B. First, we discuss how to form the new graph and structural equations in $\mathcal{M}'$. This will be necessary in cases where the antecedent or consequent are initially thought independent, like the sundowners case, or in cases where a new causal model is communicated with the conditional information. However, in many cases, this step will be unnecessary: in the disease testing example, the causal structure (where the disease causes a positive test result) remains constant through conditional learning. We focus on the case where the antecedent and consequent are initially thought independent.

Consider the case where one learns a conditional 'If A, then B' where A and B are initially thought to be independent. This means that the value of A has no effect on B: $\Pr(B|A) = \Pr(B)$. In a causal graph, we can represent this with no arrows between A and B:

$$A \qquad B.$$

Absent further information about the model, both variables would be exogenous with structural equations $A = U_A$ and $B = U_B$. This causal model can never make the conditional $A \rightarrow B$ true in a world where the consequent is not true using the conditional truth conditions from Sect. 5. Any intervention to produce A, which amounts to setting the exogenous variable U_A governing A true, would have no effect on B. This problem also arises when there is a collider in the causal graph:



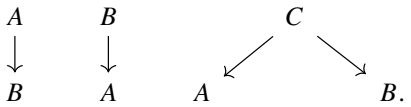
Here, since A and B are both parent nodes with no parents in common, changes in A or B are independent of each other.[10]

The problem that arises at this point is that we need a procedure which replaces the old graph, where A and B are independent, with a new one where they are conditionally dependent.[11] However, there are multiple options for how to form a new graph: having

[10] An exception to this is if C is known or held fixed, in which case A and B can become dependent conditional on C.

[11] Note that while conditional dependence is a necessary condition for the new graph to satisfy in order to learn a conditional $A \rightarrow B$, it is not always sufficient. If A is a disjunction of variable assignments, for example, we may require that the consequent is conditionally dependent on both disjuncts.

A depend on B, B depend on A, and A and B both depend on some common cause C can all render A and B probabilistically dependent:[12]



One way of trying to solve this is to find a procedure which transforms a given causal graph without conditional dependence between the antecedent and consequent into a unique new causal graph with conditional dependence. Fisher (2017b), for example, defines a measure of 'minimal illegality' for new causal models; the goal, then, would be to update the old causal model to the unique new, minimally illegal causal model.[13] However, it seems impossible to figure out whether the new causal model should have A depend on B or B depend on A just from the original causal model where two variables are independent. In the sundowners example, it is clear that rain would cause the cancelation of sundowners rather than the other way around, but this isn't encoded anywhere in the initial causal representation of the epistemic state.

Another approach is to derive a new causal model from certain syntactic or structural properties of the conditional. While providing a complete theory of the relationship between properties of conditionals and causal models goes beyond the scope of this paper, we can note some regularities in how the new causal model is formed. Conditionals with the future tense 'will' in the consequent are often causal conditionals, where the antecedent causes the consequent (i.e., 'If it rains tomorrow, sundowners will be canceled'). Conditionals where the consequent temporally precedes the antecedent, on the other hand, are usually diagnostic or 'backtracking' conditionals. For example, in the case of 'If the test comes back positive, the disease is present,' the presence of the disease temporally precedes the positive test result.

We assume that the agent learns not just a new causal ordering of the variables, but also the structural equations specifying the precise causal relationships. For causal conditionals, one learns that the antecedent is sufficient for the consequent, while with diagnostic conditionals, one learns that antecedent is necessary for the consequent. We can illustrate the difference in the corresponding structural equations by considering the sundowners example. When learning the causal conditional 'If it rains, sundowners will be canceled,' one learns not just that rain causes the cancelation of sundowners, but the full structural equations $R = U_R$ and $S = \neg R \wedge U_S$. The structural equation $S = \neg R \wedge U_S$ tells us not just that rain causes the cancelation of sundowners, but that sundowners occurs precisely when the other conditions for holding sundowners are met (U_S) and rain is absent ($\neg R$); in other words, rain is a sufficient, but not necessary, condition for cancelation. Now consider learning the diagnostic conditional 'If sundowners is canceled, it is raining.' In this case, we still have rain causing the cancelation of sundowners, but the structural equation we learn tells us that rain

[12] More generally, any graph where A and B are not d-separated can render A and B conditionally dependent. On d-separation, see Pearl (2009, p. 16).

[13] Fisher's specific proposal considers only changes to the structural equations of the model rather than the causal graph itself, so it is unclear what predictions it would offer for cases like the sundowners example.

is necessary rather than sufficient for the cancelation of sundowners, meaning the structural equation is $S = \neg R \vee U_S$: sundowners always occurs absent rain and when certain conditions (U_S) are met during rain. While further work is necessary to understand how the properties of conditionals can translate into causal constraints, we proceed under the assumption that the new conditional information is sufficient for the agent to form a new causal model.[14]

Once we have the new causal graph and structural equations for the causal model $\mathcal{M}'$ where the antecedent and the consequent are causally related, we must find the restricted set of possible worlds $U' \subseteq U$ which makes the conditional true in every world. Once we have this new U', we can update the probability distribution Pr over U to Pr$'$ by Bayesian updating, so $\Pr'(u) = 0$ if $u \notin U'$ and $\Pr'(u) = \frac{\Pr(u)}{\Pr(U')}$ if $u \in U'$. This is analogous to the procedure for learning a non-conditional proposition A in Sect. 2: we find the largest set of possible worlds U_A such that A is true in every world then update our probability distribution using Bayesian updating.

In traditional Bayesian updating, this set U_A is just the set of worlds in which A is true. For conditional learning, this is not always the case: in the disease testing example, we saw that the set of worlds in which the conditional $T \rightarrow D$ is true is the set of worlds where D is true, but this does not correspond to the content we seem to learn from the conditional. The difference is that, when we restrict to a new set of possible worlds U', we also update the set of possibilities included in the causal model $\mathcal{M}' = (U', V, f_i')$, which influences the 'set of relevant possible worlds' included in the selection function $f(A, u)$ which enters into the conditional truth conditions. Thus, to find U', we want to consider all $\widetilde{U} \subseteq U$ such that $\forall u \in \widetilde{U}$, $A \rightarrow B$ is true in u with causal model $(\widetilde{U}, V, f_i')$.

Before finding the largest such $\widetilde{U}$, there is another condition we must be attentive to for conditional learning: ensuring that we do not learn the conditional 'trivially' by just learning that the consequent is true. Consider again the case where we learn the conditional 'If the test is positive, the disease is present,' $T \rightarrow D$, where $D = U_D$ and $T = D \vee U_T$. As motivated in Sect. 3, we expect to learn the constraint that $U_T = 0$, that there are no false positives. If we find the maximal $\widetilde{U}$ such that $T \rightarrow D$ is true in $(\widetilde{U}, V, f_i')$, this will be the set of all worlds where $U_T = 0$ or $U_D = 1$. However, we do not want to include the additional worlds where $U_D = 1$; these are the worlds where we learn that the disease is present as a trivial way to make the conditional true. If we include these worlds in the new epistemic state, then the probability of the consequent will increase: using the probability distribution from Sect. 4, learning the conditional in this way would increase the probability of having the disease from $\frac{1}{10}$ to $\frac{1}{9}$.

We can eliminate these worlds by imposing a non-triviality constraint: we require that, in the model where all possible worlds are included, (U, V, f_i'), the counterfactual 'If D were not true, then $T \rightarrow D$ would still be true' must be true. This eliminates all the worlds where we make the conditional true only by setting D true. We see, for example, that the world $(U_D, U_T) = (1, 1)$ does not satisfy the condition $\neg D > (T \rightarrow D)$:

[14] There is some relevant work on this issue in cognitive science: see Griffiths and Tenenbaum (2009) and Lake et al. (2018).

intervening to set $\neg D$ sets $U_D = 0$ and the intervened world $(U_D, U_T) = (0, 1)$ is no longer a world where the conditional $T \rightarrow D$ is true (since T is true but D is false).

To return to the general case, suppose we have a causal model $\mathcal{M} = (U, V, f_i)$ and we are learning the conditional $A \rightarrow B$, where we have already found structural equations f_i' such that A and B satisfy the desired dependency relation. To update the set of possible worlds U to a new set U', we find the largest $\widetilde{U}$ such that $\forall u \in \widetilde{U}$, (1) $A \rightarrow B$ is true in u with causal model $(\widetilde{U}, V, f_i')$ and (2) $\neg B > (A \rightarrow B)$ is true in u with causal model (U, V, f_i'). With U' the largest such $\widetilde{U}$, the new causal model is given by $\mathcal{M}' = (U', V, f_i')$. If the initial epistemic state is represented by a probability distribution Pr over U, the new epistemic state consists of the Bayesian updated distribution $\Pr'$ over U'.

To see that there is a unique largest set U' satisfying these conditions, suppose that U_1 and U_2 are two distinct maximal sets. Since U_1 and U_2 are distinct, there is a $u_1 \in U_1$ which is not in U_2. This means that u_1 satisfies condition (2) and that u_1 satisfies condition (1) for (U_1, V, f_i') but not (U_2, V, f_i'). Since u_1 does not satisfy (1) in (U_2, V, f_i'), there must be some $u_2 \in U_2$ which is in the selection function $f(A, u_1)$ in U_2 but not in U_1 which prevents the conditional $A \rightarrow B$ from being true at u_1 relative to U_2. But since $u_2 \in f(A, u_1)$, $u_2 \in [A]$, and since u_2 prevents $A \rightarrow B$ from being true in u_1, $u_2 \notin [B]$. But then u_2 is a world where $A \rightarrow B$ is false in U_2 since A is true but B is false, so u_2 is a world in U_2 which does not satisfy (1). This shows that there cannot be two distinct maximal sets satisfying these two conditions, so the new set of worlds is unique.

Now that we have presented the procedure for learning conditionals, we can return to our two motivating examples to show that this procedure predicts the intuitive results motivated in Sect. 3. Consider the sundowners example, where the initial epistemic state predicts that rain R and sundowners S are independent with probability 0.5, so $R = U_R$, $S = U_S$, and $\Pr(U_R) = \Pr(U_S) = 0.5$. We learn the conditional 'If it rains, sundowners will be canceled,' $R \rightarrow \neg S$. We update our epistemic state to have a new causal structure where R causes $\neg S$ with structural equations $R = U_R$ and $S = \neg R \wedge U_S$. In this new causal model, $R \rightarrow \neg S$ is true in every world since setting $U_R = 1$ always sets $S = 0$, so we do not need to change the set of possible worlds or the distribution Pr. However, the new causal model means that the same distribution over exogenous variables induces a new distribution over endogenous variables: while $\Pr(R)$ is still $\frac{1}{2}$, S is now only true in one world, $(U_R, U_S) = (0, 1)$, so $\Pr(S) = \frac{1}{4}$. Therefore, the new epistemic state predicts that rain occurs with the same probability, but sundowners is now less likely to occur, as expected. This result generalizes for all causal conditionals: whenever we learn a causal conditional, the probability of the antecedent stays the same.

Now consider the disease testing example. The initial epistemic state predicts that the disease D causes a positive test result T, but T can occur with false positives U_T. The initial causal model specifies that $D = U_D$ and $T = D \vee U_T$, where U_D and U_T are independent and $\Pr(U_D) = \Pr(U_T) = 0.1$. We learn the conditional 'If the test is positive, then the disease is present,' $T \rightarrow D$. Since D and T are already probabilistically dependent, we do not need to change the structural equations. The largest set of worlds which makes the conditional $T \rightarrow D$ true is the set of worlds

where $U_T = 0$ or $U_D = 1$, and only those where $U_T = 0$ satisfy constraint (2). The only relevant worlds now are $(U_D, U_T) = (0, 0)$ and $(U_D, U_T) = (1, 0)$, where the former has probability 0.9 and the latter has probability 0.1 by Bayesian updating. Therefore, the new distribution predicts that $\Pr(D) = \Pr(T) = 0.1$, as expected. This result generalizes for diagnostic conditionals: learning a diagnostic conditional decreases the probability the antecedent is true.

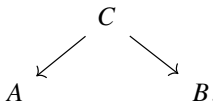
7 Further examples and extensions

Throughout, we have considered conditional learning through the lens of two examples: the sundowners problem and the disease testing case. While these examples are designed to be representative cases of causal and diagnostic conditionals, respectively, it is natural to wonder how many conditional assertions can be incorporated into the causal framework. Some more complicated examples include: (1) conditionals with no causal connection between the antecedent and the consequent, (2) conditionals with more complex causal structures than simple causal or diagnostic conditionals, (3) conditionals learned alongside other propositions, and (4) probabilistic conditionals.

One assumption of the conditional learning framework is that there is a causal connection between the antecedent and the consequent of a conditional. When this assumption is not satisfied, the truth conditions from Sect. 4 become trivial (where 'If A, then B' is equivalent to B) and the conditional learning problem cannot be approached through causal learning. While there are examples of conditionals with no causal connection between the antecedent and the consequent, these are often atypical conditionals like biscuit conditionals ('If you're hungry, there are biscuits') and Dutchman conditionals ('If he's right, then I'm a Dutchman'). Many authors think that there is something pragmatically or semantically inappropriate about conditionals where the antecedent is irrelevant for the consequent (Douven 2008; Krzyżanowska and Douven 2018), a claim which is supported by experimental evidence (Over et al. 2007; Douven et al. 2018; Skovgaard-Olsen et al. 2016). Thus, the exclusion of conditionals where the antecedent is irrelevant for the consequent does not represent a serious limitation of the causal framework.

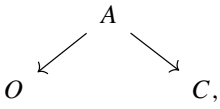
When we restrict to the conditionals where A is relevant for B, so $\Pr(B|A) > \Pr(A)$, the Markov property from Sect. 4 guarantees that the conditional relationship can be cast in causal terms. When the causal model is Markovian, whenever A is statistically relevant for B, there must be a causal relationship between A and B (Pearl 2009, p. 30). More specifically, when A is statistically relevant for B, either A causes B, B causes A, or A and B share a common cause. This suggests that, consistent with the examples of the paper, causal and diagnostic conditionals actually cover a very wide range of conditional assertions once irrelevant conditionals are excluded.

For completeness, we can briefly discuss the case of the 'common cause' conditional. A common cause conditional is a conditional 'If A, then B,' where A and B share a common cause C:



In this case, when we learn $A \rightarrow B$, we learn that A has no cause other than C and that whenever C is true, B is true. This means that the probability of A decreases since some causes of A are ruled out, as in the case of diagnostic conditionals, and that the probability of B increases since a cause of B is identified, as in the case of causal conditionals. Consider a conditional like 'If Susan is out of office next week, then she needs someone to watch her cat.' In this case, there is some common cause (e.g., a vacation) which leads Susan to be both out of office and to need someone to watch her cat. This conditional rules out causes of being out of office which do not also require a cat sitter (like a 'stay-cation'), decreasing the likelihood of being out of office. Additionally, since the conditional identifies a set of situations where a cat sitter is definitely necessary, the likelihood that a cat sitter is needed increases.

To see this illustrated more formally, suppose O represents being out of office, C represents the need for someone to watch the cat, and A represents the common cause of both, being away from home. Suppose one initially has the correct causal structure in mind,



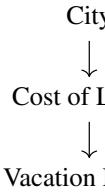
with structural equations $A = U_A$, $O = A \vee U_O$, $C = (A \wedge \neg U'_C) \vee U_C$. Here, whether Susan is away (A) is exogenous, Susan is out of office when she is away (A) or some other cause leads her to be out of office (U_O), and Susan needs someone to watch her cat when she is away (A) and she has not already made arrangements for the cat ($\neg U'_C$), or when some other cause requires someone to watch the cat (U_C).[15] To make the conditional $O \rightarrow C$ true, we eliminate the worlds where Susan is out of office but not away (U_O is activated) or Susan is away but does not need anyone to watch her cat (U'_C is activated); all of these worlds either make the conditional false or true trivially, and are therefore eliminated.[16] This redistributes probability to keep $\Pr(A)$ constant, decrease $\Pr(O)$, and increase $\Pr(C)$, as expected.

The examples in this paper are also very simple conditionals. The variables have all been binary, the antecedents and the consequents correspond to single variables rather than a conjunction or disjunction of variables, the causal structures are relatively simple, and there is no information learned alongside the conditional. However, the

[15] Here, U'_C is a disabling condition which prevents A from causing C. Such disabling conditions are common in causal models; see Pearl (2009, p. 29).

[16] If we represent worlds as (U_A, U_O, U_C, U'_C) and use $*$ to refer to a value of either 0 or 1, we can see that $(0, 1, 0, *)$, $(*, 0, 0, 1)$, and $(1, 1, 0, 1)$ are eliminated for making the conditional false and $(*, 1, 1, *)$, $(*, 0, 1, 1)$ and $(1, 1, 0, 0)$ are eliminated for making the conditional trivially true. This leaves the four worlds where U_O and U'_C are both unactivated: $(*, 0, *, 0)$.

causal framework can also handle cases of more complicated variables, logically complex conditionals, and longer causal chains. Consider the conditional 'If John moves to New York or San Francisco, he will not be able to afford an expensive vacation.' This involves variables which take on more than two values (i.e., city of residence), a disjunctive antecedent, and a longer causal chain:



However, we can still interpret this as a causal conditional with the same predictions: we learn nothing about the likely city of residence, but consider it less likely that John will be able to afford an expensive vacation since this has been ruled out if he moves to a city with a high cost of living.

The effects of learning other propositions alongside a conditional requires further discussion, especially given its prominence in recent literature (Eva et al. 2020). Following an example from Douven (2012), suppose your friend Sue recently had an important exam and you learn the conditional 'If Sue passed her exam, her father will take her on a ski vacation.' This is a causal conditional, where Sue passing the exam has a causal effect on her father taking her on a ski vacation. Following the argument of this paper, we expect learning the conditional to keep the probability of her passing the same and to raise the probability that she will go on a ski vacation, a reasonable prediction.

However, we can consider learning the same conditional with a different background, as in Douven (2012). Suppose you see Sue buying a ski outfit, so that you find it likely she is going on a ski vacation soon. Then, suppose you learn the conditional 'If Sue passed her exam, her father will take her on a ski vacation.' Given that Sue is preparing for a ski trip, it seems that learning the conditional leads one to believe it is more likely that Sue passed her exam. Since this diverges from the prediction that learning a causal conditional keeps the probability of the antecedent the same, this case calls for an explanation.

We can explain this as an effect of the background information that Sue bought a ski outfit. When one learns the conditional, one learns that the structural equation governing whether Sue goes on a ski trip is $S = P \vee U_S$, where Sue goes on a ski trip (S) when she either passes the exam (P) or she goes on a ski trip for any other reason (U_S). Updating by the information in favor of her going on the ski trip (S) then raises the probability of both P and U_S, confirming the intuition that the probability of Sue passing the exam increases. For example, if we learn with certainty that Sue is going on a ski trip, S, and we previously suspected a 10% chance of her passing and 10% chance of her going on a ski vacation without passing, our new credences upon learning that Sue is going on a ski trip (following the Bayesian updating procedure in a causal model defined in Sect. 4) would ascribe a 50% chance to Sue passing the exam. This

requires interpreting the original evidence for S retrospectively, considering it in light of the conditional information, consistent with Douven's hypothesis that explanatory considerations are relevant.[17]

Following a different example from Douven, we can also change the background so that the probability of the antecedent decreases upon learning the conditional information. Suppose we know that Sue has made plans to babysit for one of her neighbors for the whole ski season, so that it is very unlikely that she is going on a ski trip. In this case, learning the conditional 'If Sue passed her exam, her father will take her on a ski vacation' decreases the probability of the antecedent. This is because decreasing the probability of S decreases the probability of all S worlds, including the P worlds. For example, if we are certain Sue is not going on a Ski trip and learn the conditional information, the posterior likelihood that she passed the exam becomes zero.

The cases from Douven highlight how background information can change the result of conditional learning: depending on the background, learning the same causal conditional can increase, decrease, or leave unchanged the antecedent probability. While much more needs to be said to offer a complete account of how background information factors into conditional learning, we can see how Douven's predictions can be incorporated within the causal framework. We can also see how Douven's examples differ from those in this paper: while his examples rely on differences in background information to drive differences in posterior beliefs from learning a conditional, the examples in this paper predict that differences can also arise from the variations in causal structure, even when no other background information is relevant.

Another important case of conditional learning not discussed in this paper involves conditional probability constraints, such as 'If A, then the probability of B is p.' This case is particularly important because many authors prefer to model conditionals using probabilities rather than truth conditions (Adams 1975; Edgington 1995; Over and Cruz 2018). Furthermore, using probabilistic constraints may be more accurate in certain circumstances than the universal quantifier approach taken in Sect. 5. For example, in the disease testing case, very few medical tests have absolutely no false positives, and it may be more likely to encounter a constraint like 'If the test is positive, the likelihood the disease is present is 95%.'[18] While some recent accounts of conditional learning incorporate probabilistic constraints (Günther 2018; Eva et al. 2020), the differences between bare conditionals and probabilistic conditionals are significant. For example, for the bare conditional 'If the test is positive, the disease is present,' the causal framework offers a clear interpretation of the meaning as ruling out false positives. However, for the probabilistic conditional 'If the test is positive, the likelihood the disease is present is 95%,' the probabilistic constraint could arise from a decrease in the rate of false positives, a decrease in the rate of false negatives, or an increase in the prevalence of the disease, and unlike in the bare case, the meaning of the

[17] Note that this retrospective learning is essential: if one updates by S and then updates by $E \rightarrow S$ mechanistically, $E \rightarrow S$ would be trivially true and would communicate no new information. This presents an interesting case where the order of updating beliefs matters: revising by $E \rightarrow S$ and then S differs from revising by S and then $E \rightarrow S$. This issue of commutativity arises for Jeffrey conditioning (Diaconis and Zabell 1982; Wagner 2002) and poses many interesting questions, but falls outside the scope of this paper.

[18] I would like to thank an anonymous reviewer for drawing attention to the significance of conditional probability constraints.

conditional does not seem to convey which factors must change to meet the constraint. One could stipulate that all relevant factors change in the way which minimizes the Kullback–Leibler divergence (Jeffrey 1990; Van Fraassen 1981) or prioritize changing the factor which is relevant for the corresponding bare conditional, but any such theory makes assumptions beyond what is motivated by the causal account presented here. While the causal framework can likely contribute to models of belief revision handling conditional probability constraints, this issue is left open for future research.

8 Conclusion

This paper provides a causal model for learning conditionals which can explain why different conditionals lead to different posterior distributions over beliefs. Since conditional learning is observed to depend on the causal structure underlying the conditional, we introduce causal models into the theory of learning as a natural extension of Bayesian learning. After incorporating causal models into the theory of learning, conditional learning is taken to proceed in two steps. First, the old causal model is replaced by a new model which has the correct relationship between the antecedent and the consequent. Second, we restrict the set of worlds to the largest set of worlds which non-trivially sets the conditional true and then we update the probability distribution through Bayesian updating. This can account for the variations in conditional learning observed to depend on causal structure; in particular, it predicts that the antecedent probability remains constant when learning a causal conditional and that the antecedent probability decreases when learning a diagnostic conditional.

References

Adams, E. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, *6*, 89–94.
Adams, E. (1975). *The logic of conditionals: An application of probability to deductive logic*. Berlin: Springer.
Arlo-Costa, H. (2019). The logic of conditionals. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford: Metaphysics Research Laboratory, Stanford University.
Bradley, R. (2002). Indicative conditionals. *Erkenntnis*, *56*(3), 345–378.
Bradley, R. (2005). Radical probabilism and Bayesian conditioning. *Philosophy of Science*, *72*(2), 342–364.
Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*(1), 139–166.
Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association*, *77*(380), 822–830.
Douven, I. (2008). The evidential support theory of conditionals. *Synthese*, *164*(1), 19–44.
Douven, I. (2012). Learning conditional information. *Mind and Language*, *27*(3), 239–263.
Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, *101*, 50–81.
Douven, I., & Romeijn, J.-W. (2011). A new resolution of the Judy Benjamin problem. *Mind*, *120*(479), 637–670.
Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.
Eva, B., Hartmann, S., & Rad, S. R. (2020). Learning from conditionals. *Mind*, *129*(514), 461–508.
Fisher, T. (2017). Causal counterfactuals are not interventionist counterfactuals. *Synthese*, *194*(12), 4935–4957.
Fisher, T. (2017). Counterlegal dependence and causation's arrows: Causal models for backtrackers and counterlegals. *Synthese*, *194*(12), 4983–5003.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.
Günther, M. (2017). Learning conditional and causal information by Jeffrey imaging on Stalnaker conditionals. *Organon F*, *24*(4), 456–486.
Günther, M. (2018). Learning conditional information by Jeffrey imaging on Stalnaker conditionals. *Journal of Philosophical Logic*, *47*(5), 851–876.
Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In*Causal learning: Psychology, philosophy and computation* (pp. 86–100).
Hartmann, S., & Sprenger, J. (2010). Bayesian epistemology. In *Routledge companion to epistemology* (pp. 609–620).
Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.
Huisman, L. M. (2017). Learning from simple indicative conditionals. *Erkenntnis*, *82*(3), 583–601.
Jeffrey, R. C. (1990). *The logic of decision*. Chicago: University of Chicago Press.
Kaufmann, S. (2001). *Aspects of the meaning and use of conditionals*. Ph.D. Thesis, Stanford University.
Khoo, J. (2015). On indicative and subjunctive conditionals. *Philosopher's Imprint*, *15*(32), 1–40.
Kratzer, A. (1986). Conditionals. *Chicago Linguistics Society*, *22*(2), 1–15.
Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives* (Vol. 36). Oxford: Oxford University Press.
Krzyżanowska, K., & Douven, I. (2018). Missing-link conditionals: Pragmatically infelicitous or semantically defective? *Intercultural Pragmatics*, *15*(2), 191–211.
Lake, B. M., Lawrence, N. D., & Tenenbaum, J. B. (2018). The emergence of organizing structure in conceptual representation. *Cognitive Science*, *42*, 809–832.
Lewis, D. (2013). *Counterfactuals*. Hoboken: Wiley.
Nute, D., & Cross, C. B. (2001). Conditional logic. In *Handbook of philosophical logic* (pp. 1–98). Springer.
Oaksford, M., Chater, N., et al. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge internation handbook of thinking and reasoning* (pp. 434–450). Abingdon: Routledge/Taylor & Francis Group.
Over, D. E., Hadjichristidis, C., Evans, J. S. B., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.
Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.
Pollock, J. L. (1981). A refined theory of counterfactuals. *Journal of Philosophical Logic*, *10*, 239–266.
Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.
Rothschild, D. (2014). Capturing the relationship between conditionals and conditional probability with a trivalent semantics. *Journal of Applied Non-classical Logics*, *24*(1–2), 144–152.
Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, *150*, 26–36.
Stalnaker, R. (1968). A theory of conditionals. In *Ifs* (pp. 41–55). Springer.
Van Fraassen, B. C. (1981). A problem for relative information minimizers in probability kinematics. *The British Journal for the Philosophy of Science*, *32*(4), 375–379.
van Rooij, R., & Schulz, K. (2019). Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.
Vandenburgh, J. (2020). Causal models and the logic of counterfactuals. https://philpapers.org/rec/VANCMA-6.
Wagner, C. G. (2002). Probability kinematics and commutativity. *Philosophy of Science*, *69*(2), 266–278.
Weatherson, B. (2001). Indicative and subjunctive conditionals. *The Philosophical Quarterly*, *51*(203), 200–216.