Causal Inference, is just Inference: A beautifully simple idea that not everyone accepts

David Rohde d.rohde@criteo.com

Abstract

It is often argued that causal inference is a step that follows probabilistic estimation in a two step procedure, with a separate statistical estimation and causal inference step and each step is governed by its own principles. We have argued to the contrary that Bayesian decision theory is perfectly adequate to do causal inference in a single step using nothing more than Bayesian conditioning. If true this formulation greatly simplifies causal inference. We outline this beautifully simple idea and discuss why some object to it.

1 Introduction

Causal inference is often viewed as its own domain requiring concepts beyond standard probability and Bayesian decision theory. We think this complicated view is unnecessary. Bayesian decision theory automatically covers causal inference as a special case. Causal inference is complicated, not because new principles are needed but because probabilistic modelling in causal settings is difficult. Here we will show how simple Bayesian conditioning is sufficient to do causal inference and discuss why not everyone accepts the argument.

2 Bayesian Inference on Exchangeable observations

Imagine we measure an outcome on unit i, with binary outcome Y_i that received a binary treatment T_i. Furthermore, assume we have access to a dataset consisting of N different units i.e. our dataset is $Y_{1:N}$ and $T_{1:N}$. Furthermore we would like to set some future treatment T^* on another unit in the future. Our goal is to set T^* so that it will influence the outcome of Y^* and by convention we consider the outcome $Y^* = 1$ to be preferable to $Y^* = 0$. In other words the goal of our decision making problem is to determine how the treatment T^* influences the outcome Y^* and to set the treatment to maximize the probability that $Y^* = 1$.

We argue that the completely general algorithm to compute this probability is rather simple. To determine if we wish to treat $T^* = 1$ or not treat $T^* = 0$ we must specify a probabilistic model $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$, we then condition $P(Y^*|Y_{1:N}, T_{1:N}, T^*)$. Finally we compute: $\text{best } t = \text{argmax}_{t^*} P(Y^* = 1|Y_{1:N}, T_{1:N}, T^* = t^*)$. Notably, this algorithm is a straightforward application of Bayesian Decision theory, with the introduction of no novel notations or concepts to accommodate the causal aspect. Causal inference is often viewed as complex and difficult, "causation is not correlation" is a cliche of statistics. So our claim that causal inference can be reduced to computing a (Bayesian) conditional probability may be viewed with suspicion.

The point of view we develop here argues that causal inference is indeed difficult, but not because Bayesian conditioning is insufficient but rather because the task of probabilistically modelling $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ is difficult.

This modelling task is also difficult in ways that somebody familiar with using Bayesian modelling for associations might overlook. Let's consider some typical modelling assumptions that we might apply only to the observational part of the model (which is a more familiar problem to many)

Preprint. Under review.

i.e. $P(Y_{1:N}, T_{1:N}) = P(Y_{1:N}, T_{1:N}|T^*) = \int P(Y_{1:N}, T_{1:N}, Y^*|T^*)dY^*$. Usually we will assume exchangeability (or conditional independence). This is done by introducing parameters, a general way to do this is:

$$P(Y_{1:N}, T_{1:N}) = \int P(\beta, \phi) \prod_n P(Y_n|T_n, \beta, \phi)P(T_n|\beta, \phi)d\beta d\phi \quad (1)$$

Using this model we can "fill in" missing parts of the observational data. e.g. if Y_N was missing then we could compute $P(Y_N|Y_1, .., Y_{N-1}, T_1, ..., T_N)$ but equally if T_N were missing we could compute $P(T_N|Y_1, .., Y_N, T_1, ..., T_{N-1})$. The conditional probability can be viewed as "causing you to think" - or as de Finetti puts it:

> I do not look for why THE FACT that I forsee will come about, but why I DO forsee that the fact will come about. It is no longer the facts that need causes; it is our thought that finds it convenient to imagine causal relations to explain, connect and forsee the facts. Only thus can science legitimate itself in the face of the obvious objection that our spirit can only think its thoughts, can only conceive its conceptions, can only reason its reasoning and cannot encompass anything outside itself. de Finetti (1975) [7]

The cause to think interpretation allows resolution of certain associations. For example observing Christmas cards might cause you to think it is Christmas even if they do not "cause" Christmas.

There are also more restrictive assumptions, one is the following construction based on the "regression assumption":

$$P(Y_{1:N}, T_{1:N}) = \int P(\beta)P(\phi) \prod_n P(Y_n|T_n, \beta)P(T_n|\phi)d\beta d\phi, \quad (2)$$

which introduces a further partial exchangeability assumption. According to Equation 1 pairs of Y, T may be permuted i.e. The probability remains the same if $Y_i = y_i, T_i = t_i, Y_j = y_j, T_j = t_j$ or if $Y_i = y_j, T_i = t_j, Y_j = y_i, T_j = t_i$ and all other elements are the same. Assuming exchangeability allows not only exchanging pairs but arbitrary numbers of permutations.

A further exchangeability constraint is implied by Equation 2 i.e if $T_i = T_j$ then you may permute Y_i and Y_j. One way to understand this assumption is that it is only possible to learn about the association between Y_i and T_i is by observing pairs of Y and T - semi-supervised learning based on only observing T_j without Y_j is not possible.

If we were to marginalize the model to contain only $T_{1:N}$ we have $P(T_{1:N}) = \int P(\phi) \prod_n P(T_n|\phi)d\phi$. Which assumes the elements of T_i and T_j are exchangeable.

A further important remark is that this assumption does not constrain any marginal $P(Y_i, T_i)$ but does constrain the joint over $P(Y_{1:N}, T_{1:N})$. This will become important when we address critiques of probability theory as able to solve causal inference problems.

Another possibility is:

$$P(Y_{1:N}, T_{1:N}) = \int P(\alpha)P(\lambda) \prod_n P(T_n|Y_n, \lambda)P(Y_n|\alpha)d\alpha d\lambda. \quad (3)$$

Similar to above this implies partial exchangeability i.e. if $Y_i = Y_j$ then you can permute T_i and T_j and It also implies exchangeability on the marginal $P(Y_1, .., Y_N)$

We can consider three different scenarios over $P(Y_{1:N}, T_{1:N})$:

1. A model that only assumes exchangeability over pairs of Y and T using the $P(Y_n|T_n, \phi, \beta)P(T_n|\phi, \beta)$ construction

2. A model that in addition to 1. assumes partial exchangeability of Y if T is the same using the $P(Y_n|T_n, \beta)P(T_n|\phi)$ construction
3. A model that reverses the assumptions in 2. i.e. assumes partial exchangeability of T if Y is the same using the $P(T_n|Y_n, \lambda)P(Y_n|\alpha)$ construction

It is worth noting these are different probabilistic models even if as $N \rightarrow \infty$ they all converge to the same $P(Y_{N+1}, T_{N+1}|Y_1, ..., Y_N, T_{1:N})$, the difference can be seen for example in considering if semi-supervised learning is possible. In the case of 2. Having access to measurements of T_j without the corresponding Y_j provides no information how Y_k is related to T_k and so semi-supervised learning is impossible [11] in the more general case of 1. semi-supervised learning may indeed be possible.

3 Causal Inference as Bayesian Inference

At this stage we move from predicting missing elements of $Y_{1:N}, T_{1:N}$ and return to the original causal problem of determining the treatment T^* in order to induce a preferred outcome on Y^*. This requires us to model: $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$. We need to connect the new outcome Y^* to the (to be chosen by us) treatment T^* and the observed data $Y_{1:N}, T_{1:N}$. If we base our model on Equation 1 we might arrive at:

$$P(Y_{1:N}, T_{1:N}, Y^*|T^*) = \int P(\beta, \phi)P(Y^*|T^*, \beta, \phi) \prod_n P(Y_n|T_n, \beta, \phi)P(T_n|\beta, \phi)d\beta d\phi, \quad (4)$$

which unfortunately is too general for any firm conclusion to be drawn and the details of the parameteric forms and priors matter even as $N \rightarrow \infty$. In contrast this extension of Equation 2 makes strong partial exchangeability assumptions and as a consequence allows (intersubjective) causal inference:

$$P(Y_{1:N}, T_{1:N}, Y^*|T^*) = \int P(\beta)P(\phi)P(Y^*|T^*, \beta) \prod_n P(Y_n|T_n, \beta)P(T_n|\phi)d\beta d\phi, \quad (5)$$

Intersubjectivity refers to the fact that Bayesian models that agree on exchangeability but otherwise differ can rapidly reach consensus. This is a consequence of the Bayesian law of large numbers i.e. if two Bayesians agree on exchangeability but otherwise have different priors then both will have a predictive distribution that rapidly converges to the observed frequency as $N \rightarrow \infty$.

If we adopt the assumptions in Equation 5 we then assume that if we set $T^* = t$, then Y^* is exchangeable with any Y_j if $T_j = t$. In practice this means by the Bayesian law of large numbers,that as $N \rightarrow \infty; P(Y^* = 1|Y_{1:N}, T_{1:N}, T^*) \rightarrow$ empirical average of the subset of Y_j where $T_j = t$. This is the type of assumption we usually want to make when doing causal inference and this assumption is employed and appropriate after a well executed randomized control trial.

The partial exchangeability in scenario 3. where we use the $P(T_n|Y_n, \lambda)P(Y_n|\alpha)$ representation reverses the exchangeability and results in as $N \rightarrow \infty$; $P(Y^* = 1|Y_{1:N}, T_{1:N}, T^*) \rightarrow$ empirical average of the of all Y. This is the situation where T does not cause Y, which is trivial - but usefully demonstrates the impact of different partial exchangeability relationships.

Unfortunately the assumption in Equation 5 often cannot be applied (or there is disagreement about if it can be applied) and only Equation 4 might be applied which implies no use-able partial exchangeability relationship. While Equation 4 is sufficient to make causal inference very dependent on assumptions - an alternative way to demonstrate the breakdown of any useful exchangeability result is to introduce a covariate into the model and then to discuss the impact of this covariate being hidden (an unobserved confounder). Making X the covariate the model becomes: $P(Y_{1:N}, X_{1:N}, T_{1:N}, Y^*, X^*|T^*)$ If we have:

$$P(Y_{1:N}, X_{1:N}, T_{1:N}, Y^*, X^*|T^*) = \int P(\gamma)P(\eta)P(\zeta)P(Y^*|X^*, T^*, \gamma)P(X^*|\zeta) \times \prod_n P(Y_n|X_n, T_n, \gamma)P(T_n|X_n, \eta)P(X_n|\zeta)d\gamma d\eta d\zeta, \tag{6}$$

but we only observe $Y_{1:N}, T_{1:N}$ - there is no exchangeability result that can be exploited and an intersubjective treatment effect cannot be learned - it is also reasonable to expect most individual Bayesians observing $Y_{1:N}, X_{1:N}T_{1:N}$ will not learn much about $P(Y^* = 1|Y_{1:N}, T_{1:N}, T^*)$. Introducing an unobserved variable is just one way to show how exchangeability can break down. In statistical inference unobserved parameters are introduced to produce exchangeable probability models and are occasionally referred to as an indulgence in the strict "operational subjective" theory [10]. In causality unobserved confounders are introduced with the opposite purpose to destroy exchangeability and partial exchangeability between the the observed and future outcomes, but the introduction of a latent variable could equally be viewed as an indulgence.

When the covariate X is observed there are two plausible causally relevant ways a future Y^*, X^* may partially exchange with $Y_{1:N}$,$X_{1:N}$. Which results in Simpson's paradox [17]. The first of these is shown in Equation 6 with X observed, the second is given by:

$$P(Y_{1:N}, X_{1:N}, T_{1:N}, Y^*, X^*|T^*) = \int P(\gamma)P(\varpi)P(\varrho)P(Y^*|X^*, T^*, \gamma)P(X^*|T^*, \varpi) \times \prod_n P(Y_n|X_n, T_n, \gamma)P(X_n|T_n, \varpi)P(T_n|\varrho)d\gamma d\varpi d\varrho. \tag{7}$$

In the case of Equation 6 a partial exchangeability relationship exists between Y_j and Y^* so long as $X_j = X^*$ and $T_j = T^*$. In the case of Equation 7 a different partial exchangeability relationship exists between Y_j and Y^* and X_j and X^* so long as $T_j = T^*$.

4 Conclusion

Bayesian theory uses reasonable axioms of rational behaviour to show how we can use the knowledge of observed outcomes to update beliefs about other outcomes. It does not matter in principle if these observations are free form events, repetitions of a phenomena (allowing exchangeability) or are the outcome caused by a hypothetical intervention. To argue against this would require a critique of the axiom systems (See Appendix A).

It is however the case that once exchangeability is assumed as is possible in most purely observational studies the subtleties around exchangeable and partial exchangeable relationships between records can be mostly overlooked. When we must consider the causal outcome of an intervention this subtlety cannot be avoided and the probabilistic specification may be quite subjective. In this case different researchers will make a different causal inference, which is indeed a common situation when a high quality randomized control trial is not available.

It is also the case that a separate conditional probability must be computed[1] i.e. $P(Y^*, Y_{1:N}, T_{1:N}|T^* = 0)$ and $P(Y^*, Y_{1:N}, T_{1:N}|T^* = 1)$. Probability theory is entirely satisfactory to a) make causal assumptions and b) do causal inference via conditioning.

Alternative approaches separate statistical and causal inferences into separate steps. These steps involve estimation of a joint $\hat{P}(Y, T)$ and construct a causal effect as a transform of $\hat{P}(Y, T)$.

As mentioned not everybody accepts this methodology that uses probability (and partial exchangeability) both to encode associations and causal assumptions and uses only probabilistic conditioning to do the causal inference. Instead a two step procedure is adopted involving a statistical estimation of a (frequentist) distribution e.g. $\hat{P}(Y, T)$ and a causal step that explains if it is possible to recover the causal effect from $\hat{P}(Y, T)$. We argue that reducing the Bayesian $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ to the frequentist $\hat{P}(Y, T)$ obscures the partial exchangeability probability relationships that are fundamental

[1] We have no appetite to argue with anyone who sees this as an extension of probability theory.

to causal inference and requires the introduction of non-probabilistic methods both to encode causal assumptions and do causal reasoning in lieu of the simplicity and generality of probability theory. Not everyone agrees with us, and while some researchers are enthusiastic about or formulation it is rejected by key thinkers in the causal community.

In Appendix B we discuss non-probabilistic approaches to causality that separate inference into causal and statistical steps. Appendix C responds to some of the criticism and provides key references.

Appendix

A Axiom systems for Decision Making Under Uncertainty and Causality

A.1 Decision Making Under Uncertainty

In the Bayesian approach probability is a decision theoretic primitive. Probability may be defined as the price a person is willing to pay for a (reversible) bet on the outcome of a well defined future event. By the reasonable assumption that a person who provides prices for reversible bets would want to avoid being made a sure looser (a so called dutch book) it can be shown that the axioms of probability theory follow. Modern presentations of this idea can be found in [9, 10, 21, 3].

When motivated as a decision theoretic primitive probability is a consistency constraint that requires an individual is *coherent* in their probabilistic assessments. These probabilities can be applied to free form events. A textbook example may involve a joint over *it is raining*, and *the grass is wet*, and *the sprinkler is on*.

Bayesian statistics is also often applied to statistical quantities. It is equally reasonable to apply probability as a reversible bet to atomic events such as *the sprinkler is on* and to statistical quantities such as $\sum_{n=1}^{N} y_n = S$. However it is common when applying models to large numbers of repetitions of a phenomena e.g. $Y_1, ..., Y_N$ to employ an exchangeability assumption. The celebrated de Finetti representation theorem shows that such sequences can be modelled by placing a distribution over a parameter and integrating it out.

We have argued that causal inference requires a probabilistic model of $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ which factorizes:

$$P(Y_{1:N}, T_{1:N}, Y^*|T^*) = P(Y^*|Y_{1:N}, T_{1:N}, T^*)P(Y_{1:N}, T_{1:N}).$$

Of these two parts exchangeability will uncontroversially apply to $P(Y_{1:N}, T_{1:N})$ and it will not be too difficult to propose a probabilistic model that scientists agree upon. In other words an exchangeable assumption can be made in the spirit of de Finetti [7].

On the other hand probabilistic specification of $P(Y^*|Y_{1:N}, T_{1:N}, T^*)$ is likely to be much more fraught. Specification of this except in the case of a randomized control trial is likely to have a nature much more of the *the sprinkler is on* character of the Ramsey approach [18].

Causal problems therefore involve modelling challenges but both forms of this probabilistic specification tradition are entirely legitimate and de Finetti would clearly be approving of mixing them to do causal inference. Indeed to by-pass computing conditional probabilities to learn $P(Y^*|Y_{1:N}, T_{1:N}, T^*)$ would likely violate axioms of rational behavior.

A.2 Causal Reasoning

Causal analysis as presented in [13] springs from a different set of axioms. The observed system is represented by a collection of random variables using a directed ac-cyclic graph (DAG) in order to denote causality and the order that the random variables are drawn like a program. Causality is viewed as creating a new graph with some of the connections broken or mutilated. Furthermore some of the random variables are considered to be latent. The do-calculus shows if given the original distribution over the observed variables if it is possible to transform the distribution over the observables in order to recover the distribution to be expected in the mutilated graph.

The do-calculus is an impressive piece of mathematics, but it has far narrower scope than Bayesian decision theory. It is difficult to interrogate the assumptions of the DAG where tools exist for interrogating the subjective probabilities in $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ [2]. Moreover the do-calculus can only be applied in identifiable situations and ignores statistical issues.

The insistence of the inadequacy of probability theory for causal inference seems to spring from Pearl's frequentist interpretation of probability.

[2] Points of indifference to buying and selling bets can in principle be measured [7, 10]. The DAG is in contrast usually seen to make an (unverifiable) statement about the world.

> ... probability theory deals with beliefs about an uncertain, yet static world, while causality deals with changes that occur in the world itself, (or in one's theory of such changes). More specifically, causality deals with how probability functions change in response to influences (e.g., new conditions or interventions) that originate from outside the probability space, while probability theory, even when given a fully specified joint density function on all (temporally-indexed) variables in the space, cannot tell us how that function would change under such external influences. Thus, "doing" is not reducible to "seeing", and there is no point trying to fuse the two together. (Pearl 1984) [14]

A full joint over both the observed data and the post-intervention outcome means there is no need for a probability to *change*; but the idea that the probability *changes* when an intervention is applies is behind the common two step view of causal inference where there are separated statistical and causal steps each with their own logic.

B Two Step Procedures: Statistical Inference and then Causal Inference

We have argued above that the principles of causal inference are just the principles of Bayesian inference, although modelling in causal settings has specific challenges. This contrasts with other approaches that birficate the inference problem into a statistical and causal component.

In the words of Pearl "If I am remembered for no other contribution except for insisting on the causal–statistical distinction, I would consider my scientific work worthwhile" [15].

According to the two step procedure causal effects involve a statistical step to estimate a joint distribution followed by a causal step which (if possible) transforms the estimate to the causal quantities of interest.

Returning to our original example in the first step we first do a statistical analysis of $Y_1, ..., Y_N, T_{1:N}$ which may be Bayesian resulting in $P(Y_{N+1}, T_{N+1}|Y_1, ...Y_N, T_{1:N})$ or a frequentist estimate $\hat{P}(Y, T)$. The second step uses a different "causal logic" in order to consider if:

$$\begin{aligned}
&P(Y^* = y^*|Y_{1:N}, T_{1:N}, T^* = t^*)\\
&= P(Y_{N+1} = y^*|Y_{1:N}, T_{1:N}, T_{N+1} = t^*) \approx \hat{P}(Y = y^*|T = t^*)
\end{aligned}$$

as in Equation 2 or if no such assumption can be made. There are a number of ways that this causal logic may be applied.

- In the case of the *Pearlian* approach [13] if the causal graph involves an arrow from T to Y and there are no additional unobserved confounders then applying the do calculus gives $P(Y|\text{do}(T)) = P(Y|T)$.
- In the *Dawidian* approach [5] introduces a "non-stochastic-regime-indicator' F_T which switches between the observed and interventional data. If $Y \perp\!\!\!\perp F_T|T$ where $\perp\!\!\!\perp$ represents conditional independence then the causal effect is given by the conditional probability $P(Y|T)$.
- In the *Rubinesque* approach [19] a joint distribution on the counterfactual outcomes is defined where $Y_{T=0}$ is the outcome when $T = 0$ and $Y_{T=1}$ when $T = 1$. Then subject to $Y_{T=0}, Y_{T=1} \perp\!\!\!\perp T$, then the causal effect is $P(Y|T)$.

The two step procedure is a valid way to infer causal effects under limited circumstances. However reducing to a frequentist estimate of $P(Y, T)$ and consequentially the inability to access $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ requires new non-probabilistic ways to state assumptions and new non-probabilistic mathematics. Two step procedures in general suffer from the following limitations:

- Two step procedures are complicated, they unnecessarily develop different logics that apply in causal and non-causal settings.
- Two step procedures lack generality if a transform cannot be identified of the joint density then it becomes impossible to introduce prior assumptions to make an assessment of causal

effects. A Bayesian conditional probability can always be computed, and the variance of the posterior and predictive distribution can be assessed in order to determine how informative the data was[3].

- Two step procedures do not adequately handle finite sample uncertainty.
- The causal step of two step procedures are non-probabilistic in nature and invite confusion around terms such as "condition" [16].
- Two step procedures are fundamentally non-Bayesian. The notion of $P(Y, T)$ being an external stochastic process is contrary to a purist reading of Bayesian theory as "probability does not exits" it follows $P(Y, T)$ as an external entity that can be estimated also does not exist.

We do not intend to argue that there is any inherent difficulty with two step methods when they may be applied, only that they lack generality and are conceptually overly complicated.

C Response to our critics

As mentioned in the title of this document, not everyone agrees that causal inference can be reduced to Bayesian inference.

The idea presented here [12] has been twice rejected from publication, it also has been extensively discussed on Andrew Gelman's blog [8] and in a panel [6] on theoretical aspects of Bayesian Causal Inference. Several criticisms of the idea have been made.

The criticism that we are most sympathetic with is that this idea is not original. We think the correctness of the idea is more important than its originality and our contribution is surely only to clarify an existing idea. Our original manuscript presented side by side analysis of causal questions using Pearl style Causal Graphical model that were manipulated with the do-calculus and probabilistic graphical models manipulated with only probability theory. We showed both methods gave the same results using the classic examples of Simpson's paradox [17] and the front door rule [13]. We received criticism that the probabilistic graphical model we proposed was similar to twin networks already proposed in [2] and also was similar to the Bayesian approach developed in [4] where Gibbs sampling was used to marginalize out an unobserved latent confounder.

We like these papers and agree that in them Pearl and collaborators do causal inference in a single step using only probability theory very similar to our approach. Given these papers show that probability theory is adequate for causal inference it is puzzling to see Pearl advocate forcefully that causal inference requires two step procedures with different logic applying in each step.

An anonymous commentor also argued that because we motivated fully probabilistic models using Pearl graphs we had used more than probability theory [1]. That you can encode causal assumptions using only probability theory was precisely our point, in [12] we motivated the discussion with graphs in this paper we did it from first principles.

In the panel discussion [6] all panellists except Finnian Lattimore argued in favor of two step procedures. Calling the causal step a "math(s) question". We do not feel there is any clear argument made about why the causal step is non-probabilistic or distinct from inference but we invite readers to listen to this discussion. If a joint distribution of observed and latent quantities can be transformed to causal quantity of interest using only the observed quantities is indeed a maths question, but computing a conditional probability provides the causal estimate both in cases which are identifiable and non-identifiable (in the later case prior assumptions have impact even in the large data limit). Given Philip Dawid's previous Bayesian convictions we were surprised by his apparent acceptance of two step inference and wonder if this is really his considered position. The sentiment of his talk "Causal inference is just Bayesian Decision Theory" is close to that of this paper.

Turning now to reviews of our paper [12], one of the most constructive negative reviews on our submission said:

[3] If the priors have strong influence on the causal effect, then different Bayesians are likely to agree on the value of running a high quality randomized control trial to gather good estimate even if they disagree on the current causal effect estimates.

> A main point is that this all works as long as there are no latent variables; and latent variables are commonplace in causal inference. Latent variables are really latent: we know nothing about them other than they exist and that they affect some manifest variables in the graph. So they cannot be marginalised out. In this condition, the twin network is not useful. Unfortunately one cannot just decide a prior over then and proceed as usual. The authors seem to be aware of this, but still there is just no knowledge about them and we have to face it. All the technical details (parameters, other parameters, parametrisations, ...) can induce one to think wrongly about the core of the problem: some queries are just unidentifiable. (Anonymous UAI 2019 Reviewer)

A similar sentiment was put more forcefully:

> I am quite certain the method is fundamentally flawed in the presence of confounders, but even for the simpler case of non confounding not even an attempt at proof or reflection of possible assumptions / limitations is provided. (Anonymous PGM 2020 Reviewer).

Our paper actually demonstrated the agreement of probability theory and the do-calculus in the case of the front door rule which contains an unobserved latent variable (but for which causality is identifiable). In the case were causality is unidenifiable due to unobserved confounding then no partial exchangeability relationship will exist and it will not be possible to produce intersubjective causal estimates, but in contrast to the anonymous referee it is possible to place priors on the latent variables - only the affect of these priors will persist even in the large data limit. We think this framework accurately reflects how intersubjective inference can be made for well executed randomized control trials but cannot be made from natural experiments - different people have different priors and the prior impact doesn't wash out in the large data limit. Also in this case where different priors result in different inference - they typically will agree on there being very high utility in doing a well executed randomized control trial that will reduce the uncertainty. Finally as we are simply applying the Ramsey-de Finetti-Savage theory to decision making under uncertainty our method is proven to be optimal under reasonable axioms of rational behavior [18, 7, 20].

The negative PGM review also made the following comment that we find to be more substantial:

> In other word: the perceived discrepancy lies not in the assumptions or the inference rules or the available data, but strictly in the fact that the notion of 'intervention' is not part of the axioms of probability theory, and hence it needs an external frame of reference (the causal model) to make this connection. Exactly as the proposed solution in this paper does. (Anonymous PGM 2020 Reviewer)

There is indeed a point that the core construct is in some sense unusual $P(Y_{1:N}, T_{1:N}, Y^*|T^*)$ where T^* has no distribution (as we optimise it rather than integrate over it). Having a random variable Y^* having its distribution vary dependent on action T^* is unusual (and neglected) but it is present e.g. see [9] Section 7.3, but such extensions are surely under discussed and in the case of relationships between Y^* and $Y_{1:N}, T_{1:N}$ and T^* - woefully so.

We also received multiple positive reviews and comments, neutral comments and the occasional comment we did not understand. Despite the critics we are confident that this way of formulating causal inference will gain popularity due to its simplicity, generality and correctness. Tools such as the do-calculus also can provide insight and simplifications for causal problems (or indeed random variables under partial exchangeability) but should ultimately be viewed as being implied by the more general Bayesian theory.

In closing this section we would like to give the last word to Pearl who was generous enough to comment on our work in his characteristic poetic style:

> There is comfort, I admit, for researchers to dress causal inference in traditional probabilistic vocabulary; familiar words evoke familiar tools and a sense of safe passage. From logical viewpoint, however, causality and statistics do not mix, unless one extends the meaning of "statistics" to cover the entire sphere of scientific thought. (including of course speculations about Cinderella's hair color, which can be decorated with Bayes priors.) But if the comfort of traditional vocabulary

> increases researchers ability to solve causal problems (like front door, external validity, mediation and missing data) so be it — I am all for it. Judea Pearl (2020)

We think Bayesians would naturally view the meaning of "statistics" to cover the entire sphere of scientific thought including applying exchangeability to $P(Y_{1:N}, T_{1:N})$ and more free form decision making under uncertainty specification to $P(Y^*|Y_{1:N}, T_{1:N}, T^*)$. We thank Pearl for his (qualified) support.

References

[1] Anonymous. Coment on Causal inference in AI: Expressing potential outcomes in a graphical-modeling framework that can be fit using Stan. https://statmodeling.stat.columbia.edu/2020/01/27/causal-inference-in-ai-expressing-potential-outcomes-in-a-graphical-modeling-framework-that-can-be-fit-using-stan/comment-1242298, 2019. [Online; 19/9/2021].

[2] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. 2011.

[3] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[4] David Maxwell Chickering and Judea Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1269–1276, 1996.

[5] A Philip Dawid. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424, 2000.

[6] Philip Dawid, Larry Wasserman, John Langford, Finnian Lattimore, Carlos Cinelli, and David Rohde. Does causality mean we need to go beyond Bayesian decision theory? `https://www.youtube.com/watch?v=Vehb4pYf2L4`, 2019. [Online; 19/9/2021].

[7] Bruno De Finetti. *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

[8] Andrew Gelman. Causal inference in AI: Expressing potential outcomes in a graphical-modeling framework that can be fit using Stan. https://statmodeling.stat.columbia.edu/2020/01/27/causal-inference-in-ai-expressing-potential-outcomes-in-a-graphical-modeling-framework-that-can-be-fit-using-stan/, 2019. [Online; 19/9/2021].

[9] Joseph B Kadane. *Principles of uncertainty*. Chapman and Hall/CRC, 2020.

[10] Frank Lad. *Operational subjective statistical methods: A mathematical, philosophical, and historical introduction*, volume 315. Wiley-Interscience, 1996.

[11] Julia A Lasserre, Christopher M Bishop, and Thomas P Minka. Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 87–94. IEEE, 2006.

[12] Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv preprint arXiv:1906.07125*, 2019.

[13] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[14] Judea Pearl. Bayesianism and causality, or, why i am only a half-bayesian. In *Foundations of Bayesianism*, pages 19–36. Springer, 2001.

[15] Judea Pearl. *Causality*. Cambridge university press, 2009.

[16] Judea Pearl. Myth, confusion, and science in causal analysis. 2009.

[17] Judea Pearl. Comment: understanding simpson's paradox. *The American Statistician*, 68(1):8–13, 2014.

[18] Frank P Ramsey. Truth and probability. In *Readings in formal epistemology*, pages 21–45. Springer, 2016.

[19] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[20] Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

[21] J Williamson. Richard jeffrey. subjective probability: The real thing. *PHILOSOPHIA MATHEMATICA*, 14(3):365, 2006.