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.