Thesis Proposal Review: How Hard is a Causal Inference Problem

David Johnston

November 12, 2019

1 Introduction: Consequences of Decisions

This thesis is concerned with understanding a particular kind of decision problem: we are given a set of feasible decisions and a set of observed data, we know the potential consequences these decisions may have and we know how desirable these consequences are. We wish to develop strategies for selecting decisions that are likely to lead to favourable consequences. For example, the decisions may be a set of possible medical treatments, consequences are states of health and data are from published medical trials; we also assume that some states of health are known to be more desirable than others.

This general kind of problem seems to me to be a reasonable description of a type of problem that people often face (allowing that it may be somewhat simplified). But I need not rely only on an appeal to intuition to argue that this is an important class of problem, as decision problems of this type have a long and extensive history of study: Von Neumann and Morgenstern (1944) considers the problem of choosing between consequences directly with some means of evaluating their desirability, Weirich (2016) discusses decision problems featuring decisions, consequences and desirability but no explicit consideration of data. Wald (1950) considers the problem of selecting a favourable decision given a set of data and a desirability function, though he eschews explicitly considering consequences, and Savage (1972) develops Wald's theory to also include consequences of decisions, yielding a class of decision problems very similar to those discussed here. Many of the solutions presented by these authors have "entered the water supply" - in particular, the expected utility theory of Von Neumann and Morgenstern (1944) underpins an enormous amount of the work on decision problems of any type, and the risk functionals of Wald (1950) are fundamental to much of statistics and machine learning. Even theories that reject the particulars proposed by these authors build on the foundations laid by them - in short, the type of problem studied here is widely accepted to be a very important class of problem.

This type of problem has particular practical relevance to the field of *causal inference*. A Google Scholar search for "causal inference" found, in the top five results:

- Holland (1986) and Frangakis and Rubin (2002) discuss causal inference as the project of relating *treatments* to *responses* via *observations*. If we postulate an implicit desirability of responses, we have a decision problem of the type outlined
- Morgan and Winship (2014) provide in their opening paragraph three examples of causal problems. Two of them have clear interpretations as decision problems where decisions involve funding of charter schools and engaging in or encouraging college study, while the third is perhaps more concerned with *responsibility* and *remedy*:
  - Do charter schools increase test scores?
  - Does obtaining a college degree increase an individual's labor market earnings?
  - Did the use of a butterfly ballot in some Florida counties in the 2000 presidential election cost Al Gore votes?
- Pearl (2009a) begins with four examples of causal questions. The first appears to be part of a decision problem, while the second to fourth are questions of responsibility and remedy:
  - What is the efficacy of a given drug in a given population?
  - Whether data can prove an employer guilty of hiring discrimination?
  - What fraction of past crimes could have been avoided by a given policy?
  - What was the cause of death of a given individual, in a specific incident?
- Robins et al. (2000) is again concerned with estimating responses to treatments via observations

From this informal survey we have six out of ten example problems that correspond directly to the type of decision problem studied here. While decision problems are a substantial class of causal inference problems, we find that questions of responsibility also figure prominently. It is OK that there are other interesting causal questions; the focus on decision problems is justified by the fact that decision problems are an important class of problem in general, and also a large and important class of problems within causality in particular. We do not require that they are the *only* class of problems that causal researchers may be interested in.

One key difference between CSDT and existing popular approaches to causal inference is that we stipulate that *the set of decisions is a feature of the problem*, and does not depend in any way on how we choose to analyse the problem. Existing approaches provide "standard" objects (e.g. counterfactual random variables) or operations (e.g. intervening on the value of some random variable) which, if they are to be interpreted as decisions, impose some presuppositions on

the nature of the decisions available. Even if these presuppositions correspond to very common regularities of decision problems, we take the view that such regularities should be included as assumptions rather than be part of the language used to express the problem.

This difference is illustrated by the question of *external validity*. Given a randomised controlled trial (RCT), under ideal conditions existing causal inference approaches agree that certain causal effects can be consistently estimated. However, as reported by Deaton and Cartwright (2018):

> Trials, as is widely noted, often take place in artificial environments which raises well recognized problems for extrapolation. For instance, with respect to economic development, Drèze (J. Drèze, personal communications, November 8, 2017) notes, based on extensive experience in India, that "when a foreign agency comes in with its heavy boots and deep pockets to administer a 'treatment,' whether through a local NGO or government or whatever, there tends to be a lot going on other than the treatment." There is also the suspicion that a treatment that works does so because of the presence of the 'treators,' often from abroad, and may not do so with the people who will work it in practice.

Here, Drèze is describing the problem of determining the consequences of the "treatment in practice", and why these may differ from the "causal effects of treatment in the trial" - the question of external validity is, loosely, the question of how informative the latter are about the former. The usual approach of causal inference is to determine conditions under which the latter can be estimated and then, maybe, consider some additional assumptions that might allow for the latter estimate to inform the former. CSDT inverts the priority of these questions: the question of treatment in practice is primary and the question of causal effects in the trial may be a subproblem of interest under particular conditions.

One could hypothesize that CSDT and standard approaches describe two different paths that lead to the same point for an investigator prepared to accept the same assumptions. An answer to this hypothesis will require technical details of CSDT, but even at this stage we have made progress: A) we have proposed a hypothesis, B) either an affirmative or negative answer would advance the understand of the relationship between causal inference and "practical decision problems" set out in Deaton and Cartwright (2018) and C) we have proposed a necessary condition for any theory that allows us to investigate such a question, namely that *decisions are a feature of the problem*.

Bareinboim and Pearl (2012) have claimed to have a complete solution to the problem of "[identifying] conditions under which causal information learned from experiments can be reused in a different environment where only passive observations can be collected", a claim made with more force in Pearl (2018). A complete solution to the transportability of causal information is *not* a claim of a complete solution to the problem of determining the effects of "treatment in practice" or the problem of making decisions with causal information. These

latter problems ask when causal effects are informative about the consequences of the decisions in the given problem, and this question doesn't even make sense without our insistence that *decisions are a feature of the problem.*

Key features (/aims - not all are realised yet) of CSDT are:

- Conceptual clarity:
  - CSDT separates of those aspects of a problem that are fixed by non-causal considerations (objectives, feasible decisions) and causal assumptions
- Unification and extension of existing approaches to causal inference for decision problems
  - Faithful translation from any existing approach to CSDT (including the derivation of key results)
  - Exact and approximate comparison of arbitrary causal theories
  - Quantification of the *difficulty* of a causal problem
  - Necessary conditions for key results
  - Novel approaches/assumptions for causal inference

the following seems like a reasonable point, but not sure where to put it right now

The core features of CSDT are that it is a new approach to causality that is strictly more capable of representing decision problems than existing approaches, and that it allows for novel and fundamental questions to be asked. However, a secondary feature of CSDT is that its statements can be clearly resolved to statements in the underlying theory of probability. This may also be true of some counterfactual approaches, but I don't think it is true of interventional graphical models. For example, Causal Bayesian Networks feature an elementary operation notated $P(\cdot|do(\mathsf{X}_k = a))$ where X_k is a random variable on some implicit sample space E. We can ask: what does $P(\cdot|do(\mathsf{X}_k = a))$ mean in more elementary terms? $do(\mathsf{X}_k = a)$ itself *looks* like a function, and the conventional interpretation of $\mathsf{X}_k = a$ is the preimage of a under X_k. Thus, $do()$ appears to be a function typed like a measure on $\mathcal{E}$ with the domain being the sigma algebra generated by all statements $\mathsf{X}_i = a$ for all X_i associated with some graph $\mathcal{G}$, which we will denote $\sigma(\underline{\otimes}_{i\in\mathcal{G}}\mathsf{X}_i)$. We might surmise that the "conditional probability" $P(\cdot|do(\mathsf{X}_k = \cdot))$ might then be the conditional probability on $\sigma(\underline{\otimes}_{i\in\mathcal{G}}\mathsf{X}_i)$. However, CBNs in general support models where $P(\cdot|do(\mathsf{X}_k = \cdot))$ is not equal to $P(\cdot|\mathsf{A})$ for any $\mathsf{A} \in \sigma(\underline{\otimes}_{\mathcal{G}}\mathsf{X}_i)$, so our attempt to parse this notation by "conventional reading" has failed.

In fact, the situation is even more dire: we may view $do(\mathsf{X}_k = a)$ as a relation between probability measures on E which is not, in general, functional – an interpretation compatible with the definitions in Pearl (2009b). If $do()$ were functional, we could define $P(\cdot|(\mathsf{X}_k = a)$ to be the element of $\Delta(\mathcal{E})$ related to P

by $(\mathsf{X}_k = a)$. However, because $do(\mathsf{X}_k = a)$ is not functional, "conditioning" on $do(\mathsf{X}_k = \cdot)$ is ambiguous - does $P(\cdot|do(\mathsf{X}_k = a))$ refer to the set of probability measures related to P? A distinguished member of this set? In contrast to regular conditioning, where a similar ambiguity prevails but the ambient measure guarantees that disagreement can only happen on sets of measure zero, $P(\cdot|do(\mathsf{X}_k = a))$ can under different interpretations assign different measures to the same set. Causal Bayesian Network notational conventions suggest interpretations that do not make sense, and their meaning may be ambiguous even if we dig more deeply into the matter.

2 Definitions and key notation

We use three notations for working with probability theory. The "elementary" notation makes use of regular symbolic conventions (functions, products, sums, integrals, unions etc.) along with the expectation operator $\mathbb{E}$. This is the most flexible notation which comes at the cost of being verbose and difficult to read. Secondly, we use a semi-formal string diagram notation extending the formal diagram notation for symmetric monoidal categories Selinger (2010). Objects in this diagram refer to stochastic maps, and by interpreting diagrams as symbols we can, in theory, be just as flexible as the purely symbolic approach. However, we avoid complex mixtures of symbols and diagrams elements, and fall back to symbolic representations if it is called for. Finally, we use a matrix-vector product convention that isn't particularly expressive but can compactly express some common operations.

2.1 Standard Symbols

Symbol	Meaning
$[n]$	The natural numbers $\{1, ..., n\}$
$f : a \mapsto b$	Function definition, equivalent to $f(a) := b$
Dots appearing in function arguments: $f(\cdot, \cdot, z)$	The "curried" function $(x, y) \mapsto f(x, y, z)$
Capital letters: A, B, X	sets
Script letters: $\mathcal{A}, \mathcal{B}, \mathcal{X}$	σ-algebras on the sets A, B, X respectively
Script $\mathcal{G}$	A directed acyclic graph made up of nodes V and edg
Greek letters μ, ξ, γ	Probability measures
δ_x	The Dirac delta measure: $\delta_x(A) = 1$ if $x \in A$ and 0 oth
Capital delta: $\Delta(\mathcal{E})$	The set of all probability measures on $\mathcal{E}$
Bold capitals: $\mathbf{A}$	Markov kernel $\mathbf{A} : X \times \mathcal{Y} \to [0, 1]$ (stochastic maps
Subscripted bold capitals: $\mathbf{A}_x$	The probability measure given by the curried Markov kern
$A \to \Delta(\mathcal{B})$	Markov kernel signature, treated as equivalent to $A \times \mathcal{B}$
$\mathbf{A} : x \mapsto \nu$	Markov kernel definition, equivalent to $\mathbf{A}(x, B) = \nu(B)$ f
Sans serif capitals: A, X	Measurable functions; we will also call them random va
Π_{X}	The Markov kernel associated with the function X: $\Pi_{\mathsf{X}} \equiv$
$[\mathsf{A}\|\mathsf{B}]_\nu$	The conditional probability (disintegration) of A given B
$\nu\Pi_{\mathsf{X}}$	The marginal distribution of X under ν

2.2 Probability Theory

Given a set A, a σ-algebra $\mathcal{A}$ is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$
- $B \in \mathcal{A} \implies B^C \in \mathcal{A}$
- $\mathcal{A}$ is closed under countable unions: For any countable collection $\{B_i | i \in Z \subset \mathbb{N}\}$ of elements of $\mathcal{A}$, $\cup_{i \in Z} B_i \in \mathcal{A}$

A measurable space $(A, \mathcal{A})$ is a set A along with a σ-algebra $\mathcal{A}$. Sometimes the sigma algebra will be left implicit, in which case A will just be introduced as a measurable space.

Common σ algebras For any A, $\{\emptyset, A\}$ is a σ-algebra. In particular, it is the only sigma algebra for any one element set $\{*\}$.

For countable A, the power set $\mathcal{P}(A)$ is known as the discrete σ-algebra.

Given A and a collection of subsets of $B \subset \mathcal{P}(A)$, $\sigma(B)$ is the smallest σ-algebra containing all the elements of B.

Let T be all the open subsets of $\mathbb{R}$. Then $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel σ-algebra* on the reals. This definition extends to an arbitrary topological space A with topology T.

A *standard measurable set* is a measurable set A that is isomorphic either to a discrete measurable space A or $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For any A that is a complete separable metric space, $(A, \mathcal{B}(A))$ is standard measurable.

Given a measurable space $(E, \mathcal{E})$, a map $\mu : \mathcal{E} \to [0, 1]$ is a *probability measure* if

- $\mu(E) = 1$, $\mu(\emptyset) = 0$
- Given countable collection $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Write by $\Delta(\mathcal{E})$ the set of all probability measures on $\mathcal{E}$.

Given a second measurable space $(F, \mathcal{F})$, a *stochastic map* or *Markov kernel* is a map $\mathbf{M} : E \times \mathcal{F} \to [0, 1]$ such that

- The map $\mathbf{M}(\cdot; A) : x \mapsto \mathbf{M}(x; A)$ is $\mathcal{E}$-measurable for all $A \in \mathcal{F}$
- The map $\mathbf{M}_x : A \mapsto \mathbf{M}(x; A)$ is a probability measure on F for all $x \in E$

Extending the subscript notation above, for $\mathbf{C} : X \times Y \to \Delta(\mathcal{Z})$ and $x \in X$ we will write $\mathbf{C}_x$ for the "curried" map $y \mapsto \mathbf{C}_{x,y}$.

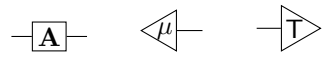
The map $x \mapsto \mathbf{M}_x$ is of type $E \to \Delta(\mathcal{F})$. We will abuse notation somewhat to write $\mathbf{M} : E \to \Delta(\mathcal{F})$, which captures the intuition that a Markov kernel maps from elements of E to probability measures on $\mathcal{F}$. Note that by similar reasoning we could consider Markov kernels to map from elements of $\mathcal{F}$ to measurable functions $E \to [0, 1]$, though we don't make use of this interpretation here.

Given an indiscrete measurable space $(\{*\}, \{\{*\}, \emptyset\})$, we identify Markov kernels $\mathbf{N} : \{*\} \to \Delta(\mathcal{E})$ with the probability measure $\mathbf{N}_*$ and there is a unique Markov kernel $\mathbf{L} : E \to \Delta(\{\{*\}, \emptyset\})$ given by $x \mapsto \delta_*$ for all $x \in E$.

We can use a notation similar to matrix-vector products to represent relationships with Markov kernels. Probability measures $\mu \in \Delta(\mathcal{X})$ can be read as row vectors, Markov kernels as matrices and measurable functions $\mathsf{T} : Y \to T$ as column vectors. Defining $\mathbf{B} : Y \to \Delta(\mathcal{Z})$ we have $\mu\mathbf{A}(G) = \int \mathbf{A}_x(G) d\mu(x)$, $\mathbf{A}\mathbf{B}_x(H) = \int \mathbf{B}_y(H) d\mathbf{A}_x(y)$ and $\mathbf{A}\mathsf{T}(x) = \int \mathsf{T}(y) d\mathbf{A}_x(y)$. The tensor product is $(mathbf A \otimes \mathbf{B})_{x,y}(G, H) = \mathbf{A}_x(G)\mathbf{B}_y(H)$ where the product on the left is scalar multiplication. Kernel products are associative and the product of kernels is always a kernel itself (Çinlar, 2011).

Some elaborate constructions are unwieldly in inline product notation. Here we use string diagrams. String diagrams can always be interpreted as a mixture of kernel products and tensor products of Markov kernels, but we introduce kernels with special notation that helps with interpreting the resulting objects. String diagrams are the subject of a coherence theorem: taking a string diagram and applying a planar deformation or any of a number of graphical rules not used here yields a string diagram that represents the same kernel (Selinger, 2010). For a thorough definition of version of string diagrams used here, see Cho and Jacobs (2019).

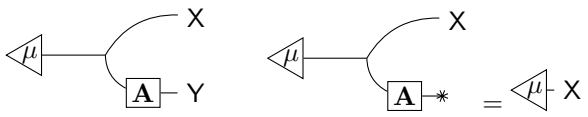
A kernel $\mathbf{A} : X \to \Delta(\mathcal{Y})$ is written as a box with input and output wires, probability measures $\mu \in \Delta(\mathcal{X})$ are written as triangles "closed on the left" and measurable functions $\mathsf{T} : Y \to T$ as triangles "closed on the right".



(1)

The identity $\mathbf{Id} : X \to \Delta(X)$ is the Markov kernel $x \mapsto \delta_x$, which we represent with a bare wire. The copy map $\curlyvee : X \to \Delta(X \times X)$ is the Markov kernel $x \mapsto \delta_{(x,x)}$. For $\mathbf{A} : X \to \Delta(Y)$ and $\mathbf{B} : X \to \Delta(Z)$, $\curlyvee(A \otimes B)_x = A_x \otimes B_x$. The discard map $*$ is the Markov kernel $X \to \{\#\}$ given by $x \mapsto \delta_\#$, where $\#$ is some one element set. Placing boxes side by side with connected wires corresponds to taking kernel products as defined above.

We will apply these notions to a couple of example constructions. Given $\mu \in \Delta(X), \mathbf{A} : X \to \Delta(Y)$ as before, the joint distribution on $X \times Y$ given by $\nu(A \times B) = \int_A A(x; B) d\mu(x)$ is given in string diagram on the left of 2. Marginalisation is accomplished with the discard map $*$; hence $\mu\curlyvee(\mathbf{Id} \otimes \mathbf{A}*) = \mu$; this is shown on the right of 2



(2)

References

Elias Bareinboim and Judea Pearl. Transportability of Causal Effects: Completeness Results. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, July 2012. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5188`.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, August 2018. ISSN 0277-9536. doi: 10.1016/j.socscimed.2017.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0277953617307359`.

Constantine E. Frangakis and Donald B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, 2002. ISSN 1541-0420. doi: 10.1111/j.0006-341X.2002.00021.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2002.00021.x`.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, December 1986. ISSN 0162-1459. doi: 10.1080/01621459.1986.10478354. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354`.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, NY, 2 edition edition, November 2014. ISBN 978-1-107-69416-3.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96–146, 2009a. ISSN 1935-7516. doi: 10.1214/09-SS057.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009b.

Judea Pearl. Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*, 2018.

James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550, September 2000. ISSN 1044-3983. URL `https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal_Structural_Models_and_Causal_Inference_in.11.aspx/`.

Leonard J. Savage. *Foundations of Statistics*. Dover Publications, New York, revised edition edition, June 1972. ISBN 978-0-486-62349-8.

Peter Selinger. A survey of graphical languages for monoidal categories. *arXiv:0908.3347 [math]*, 813:289–355, 2010. doi: 10.1007/978-3-642-12821-9_4. URL `http://arxiv.org/abs/0908.3347`. arXiv: 0908.3347.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.

Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

Paul Weirich. Causal Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL `https://plato.stanford.edu/archives/win2016/entries/decision-causal/`.

Appendix: