# AIStats Submission Sept 26: up to "CBNs are causal theories"

## 1 Introduction

It is widely accepted that causal knowledge and statistical knowledge are distinct. At least two levels are common: statistics is concerned with *association* while causation is concerned with *consequences*; a distinction of this nature goes back at least to Hume [Morris and Brown, 2019], who is also noted for his argument that knowledge of the latter cannot be reduced to the former. Pearl [2009] has identified three level hierarchy of causal knowledge in contemporary work: first *association*, then *intervention* (analogous to Cartwright's *strategy*) and finally *counterfactuals*. Pearl argues that the types of things that can be known at higher levels subsumes what can be known at lower levels (e.g. all associational knowledge is a type of interventional knowledge), but lower levels do not subsume higher levels.

An apparently paradoxical feature of this three level hierarchy is that, though knowledge is claimed to flow only in one direction, we find that the first and third levels are both described by ordinary joint probability distributions. Counterfactual queries can be formulated as missing data problems, which are distinct from associational problems only due to the interpretations we assign to so-called *counterfactual random variables* or *potential outcomes*. Knowledge at the second level, on the other hand, is described by causal graphical models which are *not* joint probability distributions (in one treatment, they are introduced as indexed sets of joint probability distributions Pearl [2009]). Here is an apparently paradoxical feature of common approaches to causal inference: associational knowledge is distinguished from consequential knowledge in both interpretation and representation, while counterfactual knowledge – considered to subsume both – is distinguished from associational knowledge by interpretation only.

Statistical decision theory, introduced by Wald [1950], underpins much of modern machine learning. It introduced the fundamental notions of *loss* and *risk* to statistics and proided foundational theorems such as the *complete class theorem* which shows that all admissible decision rules are Bayesian decision rules for some prior. Key elements of statistical learning theory inherits heavily from statistical decision theory. While some descendants of statistical decision theory have grappled with the problem of causality [Lewis, 1981], SDT itself is regarded as a theory of statistical decision making and not of causality.

Is this true? There are substantial similarities between SDT and SLT, but I haven't found direct evidence of lineage in e.g. a citation from Valiant

We show a surprising relationship between SDT and causal graphical models. We proceed in two steps: We note that a causal graphical model represents a relationship between probability measures on a given space and the consequences of a given set of actions. We then consider a modification of a standard statistical decision problem: suppose that, rather than being given a loss function that directly evaluates decisions, we are instead provided with a preference function over consequences of decisions that (following convention) we call a *utility*. The resulting problem is underspecified and provides no ordering over decisions. However, the type of relationship represented by a causal graphical model is then found to be precisely the type of object needed to fully specify the problem, and does so in a way that induces a regular statistical decision problem.

This motivates the definition of *causal statistical decision problems* (CSDPs). These relate to regular statistical decision problems (SDPs) in loose analogy with the way that model based reinforcement learning relates to model free reinforcement learning; while the former keeps track of both consequences and rewards/utililties of decisions, the latter "forgets about the consequences" and only works with reward/utiltiy.

Preprint. Under review.

CSDPs introduce the notion of *causal theories*. Causal theories represent relationships between probability measures and consequences and are a generalisation of causal graphical models. In Pearl's language, they represent the connection between associational knowledge and interventional knowledge; in Cartwright's, they connect associational knowledge with the consequences of strategies.

Thanks to the clarity of our approach, we are able to shed light on the questions raised in the second paragraph: we require a causal theory to bring knowledge from levels 1 to 2 of Pearl's hierarchy and we *also* require a causal theory to bring knowledge from level 3 to level 2. A joint distribution over counterfactuals can only answer interventional questions *given interventional assumptions* (we speculate that such assumptions may have been taken for granted). Associational knowledge is represented with probability distributions, knowledge of consequences with stochastic maps and relationships between the two with causal theories.

Choosing appropriate causal theories is a hard problem. Whether we build a causal theory with graphical models or Potential Outcomes (with additional assumptions), it is often the case that a nontrivial result rests on assumptions that are not obvious, generic or testable. Generic principles such as the bias-variance tradeoff have proved to be immensely powerful in the world of statistics, and we regard the question of whether there are generic principles that govern causal inference and what they may be to be one of the most important questions in the field.

We are primarily concerned with setting out a clear framework for reasoning about causal theories, and do not propose principles for constructing a causal theory in this paper. We are able to show a general negative result - causal theories that are symmetric over permutations of decisions cannot yield nontrivial decision rule orderings. We term this result "no causes in, no causes out" as it demonstrates that some causal knowledge is required at the outset if we hope for any nontrivial decision rules. Such asymmetric causal assumptions must be problem specific, so from the outset we cannot build causal theories on "problem neutral" assumptions alone.

There's another half baked angle here, which is "what kinds of causal theories are represented by graphical models"? In particular, via the question of dominance we can consider causal theories to be related by three different types of randomisation. Also, if we examine marginal causal models, we note that they all represent causal theories that are related to a "nice" causal theory (in the sense that identification is straightforward) via two of these types of randomisation. It's half baked because I can't yet say a lot from there, save for the fact that the operation of randomisation seems more amenable to being generalised to a continuous version than DAGs do.

I could also include the "free" results from statistical decision theory somewhere - complete class theorem, purification

## 2 Statistical Decision Problems and Causal Statistical Decision Problems

A statistical decision problem (SDP) poses the following scenario: suppose we have a set of "states of nature" $\Theta$, a set of decisions $D$ and a loss function $l : \Theta \times D \to \mathbb{R}$. For each state of nature $\theta \in \Theta$ there is an associated probability measure $\mu_\theta \in \Delta(\mathcal{E})$ where $(E, \mathcal{E})$ is some measurable space. Call the stochastic map $H : \theta \mapsto \mu_\theta$ a *statistical experiment*.Given a *decision strategy* $\pi : E \to \Delta(\mathcal{D})$, define the *risk* of $\pi$ given state $\theta$ to be the expected loss of $\pi$ in state $\theta$. Specifically, $R : \Pi \times \Theta \to \mathbb{R}$ given by $R : (\pi, \theta) \mapsto \delta_\theta \curlyvee (H\pi \otimes \mathrm{Id}_\Theta) l$, where we make use of the product notation and copy map for brevity.

Need a canonical measure on $\Theta$; the coarsest measure rendering the evaluation maps measurable?

Supposing some unknown true state $\theta^*$, we would ideally find a strategy $\pi$ that minimises the risk in $\theta^*$. Unfortunately, most statistical decision problems do not admit such strategies. Two alternative decision rules are available:

Given a measure $\xi \in \Delta(\Theta)$ called a prior, $\xi$-*Bayes decision rule* is a decision rule $\pi^*_{\mathrm{Ba}}$ such that the *Bayes risk* $R_\xi : \pi \mapsto \xi \curlyvee (H\pi \otimes \mathrm{Id}_\Theta) l$ is minimised:

$$\pi^*_{\mathrm{Ba}} \in \arg\min_{\pi \in \Pi} R_\xi(\pi) \tag{1}$$

A *minimax* decision rule $\pi^*_{\mathrm{MM}}$ minimises the worst-case risk. Unlike a Bayes rule, it does not invoke a prior:

$$\pi^*_{\mathrm{Mm}} \in \arg\min_{\pi\in\Pi} \max_{\theta\in\Theta} R(\theta,\pi) \tag{2}$$

We emphasise here that we regard the set $\Theta$ as a "state of nature" or a "theory of nature" and not a "parameter set" - it is possible that for some $\theta \neq \theta'$ we have $\mu_\theta = \mu_{\theta'}$, a possibility not supported by the interpretation of $\Theta$ as a set of distribution parameters. If there were a decision strategy that minimised the loss in every state, such a strategy would clearly minimise the loss in the true state.

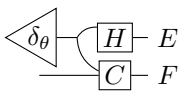
Our representation of statistical experiment is slightly different to, for example, Le Cam [1996], who introduces statistical experiments as an ordered collection of probability measures. Both representations do the same job, and the representation as a map makes for a clearer connection with causal statistical decision problems.

Formally, we define an SDP as the tuple $\langle \Theta, E, D, H, l\rangle$ where $\Theta$, $E$ and $D$ are measurable sets, $H$ is a stochastic map $\Theta \to \Delta(\mathcal{E})$ and $l$ a measurable function $E \to \mathbb{R}$. We leave implicit the set $\Pi$ of decision strategies $E \to \Delta(\mathcal{D})$ and $\mathbb{R}$, the codomain of $l$.

This is a very bare bones exposition of the theory of SDPs, and for more details we refer readers to Ferguson [1967].

Observe that a statistical decision problem supplies a loss $l$ that tells us immediately how desirable a pair $(\theta, d) \in \Theta \times D$ is. In many areas it is more typical to talk about how desirable the *consequences* of a decision are than how desirable a (state, decision) pair is. If the set of possible consequences of a decision is denoted by a set $F$, let the desirability of an element $f \in F$ be given by a utility function $u : F \to \mathbb{R}$; "utility" being a very conventional term for such a desirability function. Given such a $u$, the tuple $\langle \Theta, E, D, H, u\rangle$ is an ill-posed problem; we want to evaluate the desirability of decisions $D$ (or decision strategies $\pi$), but we have no means of connecting decisions with consequences $F$. An obvious move is to introduce for each state of nature $\theta$ a *consequence map* $\kappa_\theta : D \to \Delta(\mathcal{F})$; let $C$ be the map $\theta \mapsto \kappa_\theta$. We can then define the *causal risk* $S : \Pi \times \Theta \to \mathbb{R}$ by $S : (\pi, \theta) \mapsto -\delta_\theta \curlyvee (H\pi \otimes ID\Theta)Cu$, and Bayes and minimax risks are defined by obvious analogy.

For each state $\theta \in \Theta$, the Markov kernel

$T_\theta :=$ 

(3)

Is sufficient to compute the causal risk. Thus we can replace $H$ and $C$ with the *kernel causal theory* $T : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ given by $(\theta, d) \mapsto T_\theta(d; \cdot)$. A causal statistical decision problem (CSDP) is therefore a tuple $\langle \Theta, E, F, D, T, u\rangle$.

This is highly nonobvious, depends on the work of Jacobs, and is only known to be true for finite sets $E$ and $F$. On the other hand, there's a very intuitive graphical proof.

Given a CSDP $\alpha = \langle \Theta, E, F, D, T, u\rangle$ where $T$ is a theory arising from some $H$ and $C$ as in Equation 3, we can recover $H = T(\mathrm{Id}_E \otimes *_F)$ and $C = T(*_E \otimes \mathrm{Id}_F)$. Given $\alpha$ and letting $l := Cu$ we induce the canonical SDP $\beta = \langle \Theta, E, D, H', l\rangle$ such that for any $\theta \in \Theta$, $\pi \in \Pi$, $R^{(\beta)}(\pi, \theta) = S^{(\alpha)}(\pi, \theta)$, and thus, if we accept that the risk functional is the only means of evaluating the desirability of a strategy (whether we choose Bayes, minimax or some other meta-rule to select a strategy), $\alpha$ and $\beta$ will always produce identical recommendations.

It is also possible to induce a CSDP from an arbitrary SDP $\beta := \langle \Theta, E, D, H, l\rangle$. First, define $F := \Theta \times D$ and then let $u := -l$. Define $C : \Theta \to (D \to \Delta(\mathcal{F}))$ by $C : \theta \mapsto (d \mapsto (\theta, d))$, and then construct $T$ from $\Theta$, $H$ and $C$ as above.Then the CSDP $\alpha := \langle \Theta, E, F, D, T, u\rangle$ has the property $S^{(\alpha)}(\pi, \theta) = R^{(\beta)}(\pi, \theta)$.

Thus, in some sense every problem that can be represented as an SDP can be represented as a CSDP and vise-versa (this may not hold true if we invoke some decision rule that doesn't just depend on the risk functional).

We will finally note an alternative representation of a CSDP. Consider the set $\mathcal{T} = \{T_\theta | \theta \in \Theta\}$ and the evaulation map $\mathrm{Ev}_\mathcal{T} : \mathcal{T} \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ given by $(T_\theta, d) \mapsto T_\theta(d; \cdot)$. Then the problems $\alpha = \langle \Theta, E, F, D, T, u\rangle$ and $\alpha' = \langle \mathcal{T}, E, F, D, \mathrm{Ev}_\mathcal{T}, u\rangle$ are related in the sense that $S^{(\alpha)}(\pi, \theta) = S^{(\alpha')}(\pi, T_\theta)$. Thus, unlike with regular SDPs where $\Theta$ is an arbitrary set, for CSDPs we can regard $\Theta$ as a subset of kernels $D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ that can be written as in Eq. 3 with the

this might be equivalent to the set of 1-combs, see Jacobs

canonical kernel $\mathrm{Ev}_{\mathcal{T}}$. We call $\mathcal{T}$ the *set causal theory* and, as there is a bijection between kernel theories $T$ along with their domains $\Theta$ and set theories $\mathcal{T}$, we typically refer to either as simply a *causal theory*. We say that $\alpha$ is a CSDP in kernel form and $\alpha'$ is a CSDP in set form.

Finally, we note that causal theories can *also* be represented as a set of (distribution, consequence map) pairs.

**Theorem 2.1** (Causal theories are sets of pairs)**.** *There is a bijection between the set of causal theories and the power set* $\mathcal{P}(\Delta(\mathcal{E}) \times \Delta(\mathcal{F})^D)$

*Proof.* Sketch: Theories to pairs goes $\{(T_\theta(\mathrm{Id}_E \otimes *_F), T_\theta(*_E \otimes \mathrm{Id}_F)) | T_\theta \in \mathcal{T}\}$ and pairs to theories goes $\{\mu \otimes \kappa | (\mu, \kappa) \in \mathrm{Pairs}\}$. □

A *causal theory* - that is a "consequence-aware" analogue of a statistical experiment. It can be represented as a Markov kernel or a set of (observation, consequence) pairs. The kernel representation foregrounds the relationship between causal theories and statistical experiments, and we will exploit the latter to justify the following claim: *Causal Bayesian Networks are a subset of causal theories*.

In fact, this holds for the following generalisations of causal graphical models:

- Arbitrary sets of CBNs (e.g. ADMGs)
- Marginal models (e.g. mDAGs)
- CBNs with different intervention rules
- Cyclic SEMs (???)

don't know about this one

Show this for all the examples claimed

It is not the case, however, that *every* type of "causal graphical model" is a representation of a causal theory - SWIGs are an important exception. This is because SWIGs, like all counterfactual models, represent hypothesis classes (a hypothesis class is the range of a statistical experiment).

## 3 Causal Bayesian Networks

Suppose we have a set of "interventions" $R$ which factorises as $R = \otimes_{i \in [n]} \{\#\} \cup X^i$ for some $n \in \mathbb{N}$, collection of sets $\{X^i\}_{i \in [n]}$ and distinguished element $* \notin R^i$ for any $i$. Suppose we also have a measurable space $E$ and set of random variables $\{\mathsf{X}^i | i \in \mathbb{N}\}$ such that $\mathsf{X}^i : E \to X^i$. We denote an element $(x^0, \#, ..., \#, x^n) \in R$, $x^0, x^n \neq \#$ by the notation $do(\mathsf{X}^0 = x^0, \mathsf{X}^n = x^n)$ where occurrences of the distinguished element $*$ are ommitted. Denote by $\underline{\#}$ the element of $R$ consisting entirely of $\#$ (equivalently, $do()$).

For $n \in \mathbb{N}$, directed acyclic graph (DAG) of degree $n$ is a graph $\mathcal{G} = (V, A)$ where $V$ is a set of vertices such that $|V| = n$ and $A \subset V \times V$ is a set of directed edges ("arrows") such that $A$ induces no cycles (for a definition of cycles see Pearl [2009]).

Strictly, we are considering labeled graphs $\mathcal{G}$ and sets $\{\mathsf{X}^i\}_{[n]}$ of random variables. That is, we have bijective functions $f : V \to [n]$ and $g : \{\mathsf{X}^i\}_{[n]} \to [n]$ and we adopt the convention that $f(i) := V^i$ and $g(i) := \mathsf{X}^i$. In addition, we will sometimes let a set $U \subset V$ or $a \subset [n]$ to denote a set of random variables rather than vertices or natural numbers; this is licenced by the bijections $f$ and $g$. We will therefore overload notation and simply refer the nodes of $\mathcal{G}$ *as* the random variables $\mathsf{X}^i$.

We also suppose we have surjective $h : R \to \mathcal{P}([n])$ such that $h : (x^0, ..., x^n) \mapsto \{i | x^i \neq *\}$. That is, $h$ picks out the indices that aren't suppressed in the $do(...)$ notation for elements of $V$. Define $\mathsf{X}^{i\prime} : R \to \{\#\} \cup X^i$ by the function returning the $i$-th element of $r$ for $r \in R$. Again, we suppose we have a bijection between primed random variables and natural numbers and can therefore pick out corresponding sets of primed RVs and unprimed RVs or natural numbers.

**Definition 3.1** (Causal Bayesian Network)**.** Given $R$, $E$ and $P_* : R \to \Delta(\mathcal{E})$ and $\{\mathsf{X}^i\}_{i \in [n]}$, a Causal Bayesian Network (CBN) compatible with $P_*$ is a directed acyclic graph (DAG) $\mathcal{G}$ of degree $n$ such that for all $r \in R$

1. $P_r$ is compatible with $\mathcal{G}$ (see Pearl [2009])

2. For all $i \in h(r)$, $P_r F_{\mathsf{X}^i} = \delta_{\mathsf{X}^{i\prime}(r)} F_{\mathsf{X}^i}$

3. For all $i \notin h(r)$, $P_{r|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i} = P_{\underline{\#}|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i}$, $P_{\underline{\#}}$-almost surely

This definition differs slightly from that given in Pearl [2009]; for example $P_*$ is a map to $\Delta(\mathcal{E})$ rather than a set of labeled members of $\Delta(\mathcal{E})$, and we forumlate it in directly in terms of measure theoretic probability rather than elementary probability. Nonetheless, we claim these choices don't meaningfully alter the standard definition, at least if we restrict $E$ to be finite, and they make for a more convenient connection with CSDPs.

I haven't found a formulation of CBNs on infinite spaces, let alone continuous ones

A graph $\mathcal{G}$ and a measure $\mu \in \Delta(\mathcal{E})$ compatible with $\mathcal{G}$ together define a class of stochastic maps $K \subset \Delta(\mathcal{E})^V$ such that every $P_* \in K$ is compatible with $\mathcal{G}$ and $P_*(\#) = \mu$. Let the notation $\mathcal{G}(\mu)$ stand for the set $K$ as defined here; note that $\mathcal{G}(\mu)$ is in general a set-valued function.

At least in the case of discrete $E$ and $P_*(\#)$ positive definite, we have from this definition for any $r \in V$ the *truncated factorisation* property:

$$P_r F_{\mathbf{X}}(A) = \prod_{i \in h(r)} \delta_{\mathsf{X}^{i\prime}(r)}(\mathsf{X}^i(A)) \sum_{a \in A} \prod_{i \notin h(r)} P_{\#|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i}(a; \{\mathsf{X}^i(a)\}) \tag{4}$$

As a consequence of the existence of conditional probability, given $\mathcal{G}$ and $\mu$ there exists a unique set of interventional maps $P_*$ compatible with both $\mathcal{G}$ and $\mu$ as above. This property licenses a typical use case of CBNs: $\mathcal{G}(\cdot)$ is treated as a *map* from the subset of $\Delta(\mathcal{E})$ compatible with $\mathcal{G}$ to interventional maps $V \to \Delta(\mathcal{E})$. More generally, provided $\mu$ is compatible with $\mathcal{G}$ we have that 4 exists, and so $\mathcal{G}(\mu)$ is non-empty.

General definition: $f(A)$ is the image of $A$ under $f$ and $X$ as the copy-mapped tensor product of variables in set $X$

Condition 3 presents some difficulties in the presence of measure 0 sets, as when a conditional probability such as $P_{\#|\mathrm{Pa}_{\mathcal{G}}(\mathsf{V}^i)}$ may be variously intended to mean a particular element of the class of conditional probabilities, any element or every element in the class class [Çinlar, 2011]; condition 3 will have different implications for these various interpretations.

Letting $\mathcal{H}^{\mathcal{G}} \subset \Delta(\mathcal{E})$ be some *hypothesis class* of probability measures compatible with a causal graph $\mathcal{G}$, define the set of pairs $\mathcal{T}^{\mathcal{G}} := \{(\mu, \kappa) | \mu \in \mathcal{H}, \kappa \in \mathcal{G}(\mu)\}$. Recall that a causal theory can be represented as a set of (observation, consequence map) pairs; i.e. $\mathcal{T}^{\mathcal{G}}$ is a causal theory. The map $\mathrm{Th} : \mathcal{G} \mapsto \mathcal{T}^{\mathcal{G}}$ is therefore a map from directed acyclic graphs to causal theories. Unlike the map from DAGs to sets of probability measures, the map from DAGs to causal theories is injective.

Surely Pearl or a student has dealt with this somewhere? Any element seems to be the most appropriate choice, but this renders CBNs useless for continuous spaces unless we place extra restrictions on $P_*$

**Theorem 3.2** (The map Th is injective)**.** *For DAGs $\mathcal{G}$, $\mathcal{G}'$ on the same set of RV's $\{\mathsf{X}^i\}_{[n]}$, $\mathcal{G} \neq \mathcal{G}' \implies \mathcal{T}^{\mathcal{G}} \neq \mathcal{T}^{\mathcal{G}'}$.*

*Proof.* Sketch: $\mathcal{G}$ and $\mathcal{G}'$ must disagree on at least one parental set. Choose some $\mu$ such that $P_{\underline{\#}|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i} \neq P_{\underline{\#}|\mathrm{Pa}_{\mathcal{G}'}(\mathsf{X}^i)} F_{\mathsf{X}^i}$. Then take $r, r'$ such that $h(r) = h(r') = \mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i) \cup \mathrm{Pa}_{\mathcal{G}'}(\mathsf{X}^i)$, $\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}'^i)(r) = \mathrm{Pa}_{\mathcal{G}}(\mathsf{X}'^i)(r')$ but $r \neq r'$. Then $P_r F_{\mathsf{X}^i} \neq P_{r'} F_{\mathsf{X}^i}$ so $P_r \neq P_{r'}$. □

Each DAG $\mathcal{G}$ represents a causal theory $\mathcal{T}^{\mathcal{G}}$. For every causal theory $\mathcal{T}$, either it is not represented by any graph or there is a unique graph $\mathcal{G}$ such that $\mathcal{T} = \mathcal{T}^{\mathcal{G}}$. It is in this sense that we claim Causal Bayesian Networks are a subset of causal theories.

## 4 potential outcomes models

We will follow Rubin [2005] in our development of potential outcomes models, acknowledging that there are a wider variety of approaches to modelling potential outcomess than Rubin's version potential outcomes alone. Given an underlying state space $\Theta$, Rubin posits a number of Markov kernels and composes them to define a statistical experiment $H : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{W})$. We eschew the "conditional probability" notation Rubin uses as it masks the distinction between a Markov kernel that is the disintegration of a previously defined joint probability distribution and a Markov kernel that we are simply defining as such, and will not necessarily be a disintegration of any probability distribution.

Notationally, we will refer to the symbol $\mathsf{W}_i$ as the $i$-th treatment assignment ($i \in \{0, .., n\}$), $\mathsf{Y}_i(0)$ $\mathsf{Y}_i(1)$ as the $i$-th potential outcomes, $\mathsf{Y}_i$ as the $i$-th obseved outcome and $\mathsf{X}_i$ as the $i$-th "vector of

background facts". $\mathsf{W}$ refers to the composite of all $\mathsf{W}_i$s and similarly for other symbols. We avoid strictly defining what these symbols represent for now, as the construction of $H$ allows us to be more explicit about what exactly these symbols represent. Suppose the vector $[\mathsf{Y}_0, ..., \mathsf{Y}_n]$ takes values in $Y$ and similarly for other symbols.

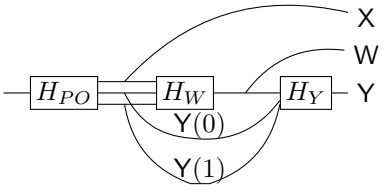
In particular, Rubin supplies:

- A "model on the science", $H_{PO} : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Y})$ (In Rubin's notation, $\prod_i f(\mathsf{X}_i, \mathsf{Y}_i(0), \mathsf{Y}_i(1)))$
- An "assignment mechanism", $H_W : X \times Y^2 \to \Delta(\{0,1\}^n)$ (in Rubin's notation, $Pr(\mathsf{W}|X, Y(1), Y(0)))$
- An "observation model", $H_Y : \{0,1\}^n \times Y^2 \to \Delta(\mathcal{Y})$, defined explicitly as $H_Y : (\mathbf{y}^0, \mathbf{y}^1, \mathbf{w}) \mapsto (1 - \mathbf{w}) \odot \delta_{\mathbf{y}^0} + \mathbf{w} \odot \delta_{\mathbf{y}^0}$ where $\odot$ is the elementwise product

Note that this construction does not permit the usual assumption of consistency ($W_i = w \implies \mathsf{Y}_i = \mathsf{Y}_i(w)$) because Markov kernels can at best give almost sure equality.

We then define the experiment $H$ by

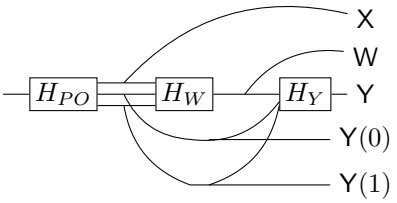
$H :=$




(5)

Where we have labeled the wires "carrying" $\mathsf{Y}(0)$ and $\mathsf{Y}(1)$ for clarity. In fact, the "vector" symbols defined above are naturally associated with wires in the above diagram. Addtionally, we could draw an alternative diagram where each wire was copied $n$ times to reflect the unit level symbols $\mathsf{X}_i$ etc. It is also possible to define the symbols as random variables, though we require an expenanded sample space $\Omega := X \times W \times Y^3$ to do so. Consider the kernel $H^*$ which is constructed from the same components as $H$:

$H^* :=$




(6)

Then $\mathsf{X}$ is the random variable identical to the projection $\pi_X : \Omega \to X$ and similarly for other symbols.

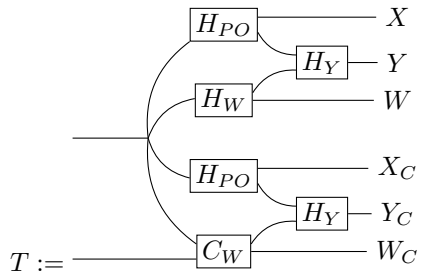
We are interested here in defining a "general type" of potential outcomes problem rather than investigate particular assumptions that may permit inference. Without a formal guide as to how to do this, we will postulate that a potential outcomes model is, in general, three Markov kernels $\langle H_{PO}, H_W, H_Y \rangle$ where $H_{PO} : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y}^m)$, $H_W : \Theta \to \Delta(\{0, .., m-1\}^n)$ and $H_Y : \{0, .., m-1\}^n \times Y^m \to \Delta(\mathcal{Y})$ which is a "selection function" in the sense defined above. We adopt the alternative signature for $H_W$ as it seems reasonable to suppose that the details of this kernel aren't always known *a priori*. Note that in general multiple potential outcomes models will yield the same statistical experiment $H$, though we postulate that in general different models will yield different $H^*$.

### 4.1 Can we consider potential outcomes models to be causal theories?

A potential outcomes model is a statistical experiment. Therefore, given a tuple $\langle \Theta, E, D, H, u \rangle$ where $H$ follows from a potential outcomes model $\langle H_{PO}, H_W, H_Y \rangle$ and $u : F \to \mathbb{R}$ is a utility

function, we have an ill-posed causal problem. A potential outcomes model is not literally a causal model, but we might ask if it is possible that our potential outcomes model $H$ is a "causal theory in disguise"; Is there a natural map from potential outcomes models to causal theories?

We will propose, somewhat weakly, that given a well-specified potential outcomes model, decisions correspond to modifications of $H_W$. As there is no general way to identify an arbitrary set of decisions $D$ with different assignment functions $H_W$, we offer (weakly) that the answer to the question in the paragraph above is "no". However, given knowledge of the "decision-influence treatment assignment" $C_W : \Theta \times D \to \{0,1\}^n$, we *can* define a causal theory via the four elements $\langle H_{PO}, H_W, H_Y, C_W \rangle$. We've supposed here there are $n$ "observational" units and $n$ "consequence" units, a restriction that simplifies the notation and is fairly easy to lift. Concretely, the causal theory is:




(7)

Where $\mathsf{X}_C$, $\mathsf{W}_C$, $\mathsf{Y}_C$ are the "consequence" analogues of observational variables $\mathsf{X}$, $\mathsf{W}$ and $\mathsf{Y}$. To simplify the diagram, we have merged the wires for $\mathsf{Y}(0)$ and $\mathsf{Y}(1)$ and ommitted the labels; the potential outcomes are carried by the wire from $H_{PO}$ to $H_Y$. Without detailed justification, we will note that this construction is unlikely to be appropriate if $H_{PO}$ does not define an exchangeable sequence of potential outcomes, an assumption that we have made as a result of following Rubin [2005].

For a "typical use" the answer is negative. Suppose a potential outcomes model $\langle H_{PO}, H_W, H_Y \rangle$ is used in the evaluation of a public program, and it is intended to inform decisions 0: cut funding or 1: maintain funding. We construct a causal theory by specifying the action of these decisions:

- Under $d_1$, state is mapped to consequences via an unchanged experiment $H$
- Under $d_0$, the state is mapped to consequences via a modified experiment $H'$ generated by $\langle H_{PO}, H'_W, H_Y \rangle$ where $H'_W : \_ \mapsto 0$

Define the causal theory $T$:

$$T : (\theta, d; A) \mapsto \begin{cases} H'(\theta; A) & d = 0 \\ H(\theta; A) & d = 1 \end{cases} \tag{8}$$

Supposing $Y = [0,1]$ and positing a utility function $u := \pi_Y$, we can compare the utilities of decisions 0 and 1 for state $\theta$ by $Tu(\theta, 1) - Tu(\theta, 0)$ (in more familiar notation, $\mathbb{E}_{T(\theta,1;\cdot)}[u] - \mathbb{E}_{T(\theta,0;\cdot)}[u]$). By construction, if we let $H^*$ be the "expanded" version of $H$ above, $Tu(\theta, 1) - Tu(\theta, 0) = H^*_{\theta|\mathsf{W}}\pi_{\mathsf{Y}(1)}(1) - H^*_{\theta|\mathsf{W}}\pi_{\mathsf{Y}(0)}(1)$; this is because only the units for which $\mathsf{W} = 1$ have different outcomes under the different decisions. This quantity is known as the *effect of treatment on the treated* (ETT) [Heckman, 1991].

We can also consider the problem in medicine of evaluating the "effect of assigning treatment" vs "the effect of receiving treatment" (the former being known as *intention to treat* analysis). From Shrier et al. [2017]:

> In public health, we are normally concerned with the first question – the effect of assigning a treatment. If we implement a prevention or treatment program that is efficacious only under strict research conditions but people in the real world would not receive it for any possible reason, the program will not be effective. This real-world context is termed the "average causal effect" of assigning treatment and is best estimated by the intention-to-treat (ITT) analysis [...]

> There are 2 reasons why the average causal effect ofreceiving a treatment may be more important than the ITT for some people. First, even in the public health domain, investigators may want to know what the average causal effect of a treatment program would be if they could improve participation in the program. [...] Also, the average causal effect of receiving a treatment is of primary interest to a patient deciding whether or not to take the treatment as recommended.

As Shrier suggests, we can consider the potential outcomes $\mathsf{Y}(0)$ and $\mathsf{Y}(1)$ to represent the outcomes of an individual who is merely *assigned* a treatment or of an individual who is *actually given* a treatment. Note that in the former case, at least for a public health decision maker, one could reasonably suppose that the assignment function $H_W$ was fully controlled by the decisions available - we can be absolutely certain, choosing $d = 1$, that the patient is assigned a treatment and likewise that they are not assigned for $d = 0$. However, if the patient's chance to actually take the treatment may differ from experimental conditions - or even worse, if we are the patient and we *know* that we will take the treatment if we should decide to - then it is unclear how the potential outcomes model given would help to determine the expected consequences of a decision. On the other hand, if we suppose that potential outcomes represent the outcomes of *taking* a treatment, then we may be able to understand the difference between the three scenarios as differences in how decisions will relate to the "treatment taking" function $H_W$ - in the first case, we know it to be the same as the experimental treatment taking function though we may not know what that function is, in the second we do not know this but we might believe it is similar and in the third case we know what $H_W$ is under either decision – deciding to take the treatment means we definitely take the treatment and vise versa.

While a potential outcomes model contains more structure than the induced statistical experiment $H$, the issue here appears to be closely related to *No Causes in No Causes Out* (Theorem **??**): potential outcomes models on their own do not appear to tell us enough about "how decisions relate to consequences". Pragmatically, at least, we should not consider a potential outcomes model to be synonymous with a causal theory.

### 4.2 potential outcomes models via Structural Equation Models

## References


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

Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, July 1967. ISBN 978-1-4832-2123-6.

James J. Heckman. Randomization and Social Policy Evaluation. SSRN Scholarly Paper ID 995151, Social Science Research Network, Rochester, NY, July 1991. URL https://papers.ssrn.com/abstract=995151.

L. Le Cam. Comparison of Experiments - A Short Review.pdf. *IMS Lecture Notes - Monograph Series*, 30, 1996. URL https://www.fastmailusercontent.com/mail-attachment/f2132276u33762/%3CB868E7FAD8F0C04FA0207113AD7C77A9%40ausprd01.prod.outlook.com%3E/Comparison%20of%20Experiments%20-%20A%20Short%20Review.pdf?u=9a2e41ab&a=195efd739077a3e70251567c111372cb67a6dcf0.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, March 1981. ISSN 0004-8402. doi: 10.1080/00048408112340011. URL https://doi.org/10.1080/00048408112340011.

William Edward Morris and Charlotte R. Brown. David Hume. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019. URL https://plato.stanford.edu/archives/sum2019/entries/hume/.

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

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

Ian Shrier, Evert Verhagen, and Steven D. Stovitz. The Intention-to-Treat Analysis Is Not Always the Conservative Approach. *The American Journal of Medicine*, 130(7):867–871, July 2017. ISSN 1555-7162. doi: 10.1016/j.amjmed.2017.03.023.

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