The Logic of Counterfactuals and the Epistemology of Causal Inference

Hanti Lin

University of California, Davis

ika@ucdavis.edu

November 30, 2024

Abstract

The 2021 Nobel Prize in Economics recognizes a type of causal model known as the Rubin causal model, or potential outcome framework, which deserves far more attention from philosophers than it currently receives. To spark philosophers' interest, I develop a dialectic connecting the Rubin causal model to the Lewis-Stalnaker debate on a logical principle of counterfactuals: Conditional Excluded Middle (CEM). I begin by playing good cop for CEM, developing a new argument in its favor—a Quine-Putnam-style indispensability argument. This argument is based on the observation that CEM seems to be indispensable to the Rubin causal model, which underpins our best scientific theory of causal inference in health and social sciences—a Nobel Prize-winning theory. Indeed, CEM has long remained a core assumption of the Rubin causal model, despite challenges from within the statistics and economics communities over twenty years ago. I then switch sides to play bad cop for CEM, undermining the indispensability argument by developing a new theory of causal inference that dispenses with CEM while preserving the successes of the original theory (thanks to a new theorem proved here). The key, somewhat surprisingly, is to integrate two approaches to causal modeling: the Rubin causal model, more familiar in health and social sciences, and the causal Bayes net, more familiar in philosophy. The good cop/bad cop dialectic is concluded with a connection to broader philosophical issues, including intertheory relations, the revisability of logic, and the role of background assumptions in justifying scientific inference.

1 Introduction

This is an invitation to the *Rubin causal model* (Rubin 1974), also known as the *potential outcome framework*. Designed for causal inference, this framework has been highly influential in the health and social sciences, underpinning the 2021 Nobel Prize in Economics. However, it is under-discussed in philosophy, and deserves far more attention than it currently receives.[1] Since philosophers crave arguments, I will try to spark their interest by developing a debate around the Rubin causal model and linking it to a familiar controversy over a logical principle of counterfactuals:

> Conditional Excluded Middle (CEM)
>
> It is logically necessary that
> - either B would be the case if A were the case,
> - or B would not the be case if A were the case.

Due to the severe lack of discussion of the Rubin causal model in philosophy, I will have to play both good cop and bad cop myself. Specifically, I will first develop a new argument for CEM. Here is the idea: CEM was already assumed in the early days of the Rubin causal model (Rubin 1974), which found important applications to causal inference through the work of Imbens & Angrist (1994), culminating in the 2021 Nobel Prize in Economics. Notably, even though the assumption of CEM was challenged in the scientific community more than 20 years ago (Dawid 2000), it has remained central to the Rubin causal model to this day. Thus, CEM appears to be an indispensable part of our best scientific theory of causal inference in health and social sciences. A new argument for CEM then emerges: an indispensability argument in the style of Quine (1948) and Putnam (1971), as detailed below (Section 3).

Next, following the good cop/bad cop approach, I will switch sides and challenge the indispensability argument. A new theory of causal inference will be developed to dispense with CEM while preserving its Nobel-Prize-winning applications. The trick, somewhat surprisingly, is to combine two causal modeling frameworks: the Rubin causal model, more familiar to health and social scientists, and the causal Bayes net, more familiar to philosophers (Section 4).

[1] For example, *the Stanford Encyclopedia of Philosophy* includes an entry on the Philosophy of Economics (Hausman 2024), which has a section dedicated to causation in economics (section 2.5) but does not mention the Rubin causal model. The entry on Causal Models (Hitchcock 2024) cites Rubin's works but does not refer to any philosophical discussions in the literature.

In the final section, 5, the good cop/bad cop dialectic will conclude by connecting it to additional issues familiar in philosophy, such as the revisability of logic in light of empirical sciences, the role of background assumptions in justifying scientific inference, and intertheory relations.

But before all that, I will have to begin by developing an accessible tutorial on the Rubin causal model for philosophers, presented in Section 2. This tutorial will be accompanied by a fully rigorous version in Appendix A, which improves upon textbook-standard presentations in important ways for philosophical purposes.

2 The Rubin Causal Model without Tears

The Rubin causal model has been extensively applied to study various aspects of our medical and economic lives. Think about it: life itself is not unlike a card game.

2.1 Introducing the Card Game

There are cards that determine our fates:

Card #1: What If You Took the Treatment?

> Nature gives every individual a card of this form: the back is printed with '*if Take* = 1', and the face is printed with '*Cure* = 1' or '*Cure* = 0'.

The former case means that this person would be cured if they took the treatment, while the latter means that this person would *not* be cured if they took the treatment. Thus, this card design already presupposes Conditional Excluded Middle.

There is only one rule for card flipping: any card given to a person is initially face down and will be flipped to reveal the result exactly when the if-clause actually applies to that person.

Similarly, there is also:

Card #2: What If You Didn't Take the Treatment?

> Nature gives every individual a second card, with the back printed '*if Take* = 0', and the face is printed with '*Cure* = 1' or '*Cure* = 0'.

Each person's cards #1 and #2 define that person's *individual treatment effect* (ITE): the value of binary variable *Cure* on card #1 minus its value on card #2.

There are three possible cases:

$$\text{ITE} \;=\; \begin{cases} 1 & (=1-0) & \text{i.e. improvement,} \\ 0 & (=1-1 \text{ or } 0-0) & \text{i.e. no difference,} \\ -1 & (=0-1) & \text{i.e. deterioration.} \end{cases}$$

The *average treatment effect* (ATE) for a population is defined as the average of the individual treatment effects for all individuals in the population.

A bit of algebra shows that the ATE is equal to the difference between two proportions:

$$\begin{aligned} \text{ATE} \;=\;& \text{(i) the proportion of `}\mathit{Cure}=1\text{' cards among all cards of type \#1} \\ & - \text{(ii) the proportion of `}\mathit{Cure}=1\text{' cards among all cards of type \#2.} \end{aligned}$$

There is a simple and effective way to estimate term (i): randomly flipping some cards of type #1 in the population—or equivalently, randomly selecting some people in the population and *forcing* them to flip their cards of type #1. Once the faces of those cards are revealed, register the proportion of the occurrences of '$\mathit{Cure}=1$', and use it as an estimate of term (i). Term (ii) can be estimated similarly. This procedure for estimating the ATE is the idea behind *randomized controlled trials* (RCTs). The problem, however, is that RCTs are often ethically impermissible.

Fortunately, there is a Nobel-Prize-winning solution, which seeks to estimate, not exactly the ATE, but a closely related causal effect—without forcing anyone to do anything.

2.2 Switching from the ATE to the LATE

Let's randomly select individuals from the population and then assign each of them to either the treatment or control group by flipping a coin. Here is the thing: anyone in the treatment group is offered the treatment for free, and they decide whether to take it—there is no forcing anyone to do anything. This creates a new type of card:

> **Card #3: What If You Were Assigned to the Treatment Group?**
>
> Nature gives every individual a card of this form: the back is printed with '*if* $\mathit{Assign}=1$' (where 1 means the treatment group), and the face is printed with '$\mathit{Take}=1$' or '$\mathit{Take}=0$'.

This determines whether the individual would or would not take the treatment if assigned to the treatment group. Similarly:

Card #4: What If You Were Assigned to the Control Group?

Nature gives every individual a card of this form: the back is printed with '*if Assign* = 0' (where 0 means the control group), and the face is printed with '*Take* = 1' or '*Take* = 0'.

With the new cards, we can define some subpopulations:

1. *Compilers*: those who would take the treatment if assigned to the treatment group, and would not if assigned to the control group (namely, those whose card #3 and card #4 are printed with '*Take* = 1' and '*Take* = 0', respectively).

2. *Defiers*: those who would do the opposite of what compliers would do.

3. *Always-Takers*: those who would take the treatment regardless of assignment.

4. *Never-Takers*: those who would not take the treatment regardless of assignment.

By Conditional Excluded Middle, those four subpopulations jointly exhaust the entire population.

Now, let the target of estimation be, not exactly the ATE, but a closely related quantity, the LATE, short for *local average treatment effect*. The LATE is defined as the average of the individual treatment effects of just the *compliers* in the population, or more formally:

$$\text{LATE} \quad =_{\text{df}} \quad \frac{\sum_{i\text{: being a complier}} \text{ITE of individual } i}{\text{the number of compliers}} .$$

Interestingly, when there are no defiers and other conditions are met, it is possible to estimate the LATE without forcing anyone to take the treatment, as will be shown shortly.

2.3 Estimating the LATE

The standard procedure for estimating the LATE is known as *instrumental variable estimation*. To understand it, we need a theorem, now a classic result in econometrics and statistics (Imbens & Angrist 1994, Angrist, Imbens, & Rubin 1996):

Informal Statement of Theorem 1 (Identification of the LATE). *In the card game presented above, which already builds in Conditional Excluded Middle, suppose that the following four assumptions hold:*

- (Random Selection) *People are randomly selected from the population.*
- (Random Assignment) *The selected people are randomly assigned to the treatment or control group.*
- (Existence of Compliers) *There are compliers.*
- (No Defiers) *There are no defiers.*

Then the LATE *can be expressed solely in terms of probabilities over the three observable variables—Assign, Take, and Cure—without counterfactuals. Specifically:*

$$\text{LATE} \;=\; \frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)}\,.$$

To be more precise, this equation holds under the assumptions 1-8 formalized in Appendix A. The first four of those assumptions, which include CEM, are automatically build into the card design.

Some explanations are in order. First, the assumption that there are compliers plays a straightforward role by ensuring that the target of estimation, the LATE, is well-defined (i.e., has a nonzero denominator).

The assumption of no defiers plays a more interesting role: to delineate the scope of application. For example, when estimating the causal effect of a newly designed drug not yet available on the market, no one in the control group could take the new drug, which implies that no one is a defier. Another example comes from Angrist's (1990) now-classic study on the Vietnam War, where "random assignment" refers to the draft lottery, "treatment" to military service, and the "medical result" to lifetime earnings. A defier in this scenario is someone being this crazy: one who would volunteer for military service if they were not drafted but would avoid service if drafted. Here, it is also reasonable to assume that no defiers exist. However, in cases where it is implausible to assume the absence of defiers, the theorem above provides no guidance on estimating causal effects.

Let's now turn to Pr, the probability function in use. The probabilities discussed in this paper are restricted to physical objective probabilities. These probabilities might be frequencies (Neyman 1955), propensities (Popper 1959), or primitive physical states posited in science (Sober 2000: sec. 3.2)—to mention just the options developed with classical statistics in mind, which often serves as the background theory for the Rubin causal model. I remain open to the metaphysics of physical objective probabilities; the focus of this paper is epistemology.

The first conditional probability in the equation, $\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1)$, is defined in the standard way:

$$\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1) \;\; = \;\; \frac{\Pr(\mathit{Cure} = 1 \wedge \mathit{Assign} = 1)}{\Pr(\mathit{Assign} = 1)}\,,$$

where the denominator is the probability that a randomly selected person is assigned to the treatment group ($\mathit{Assign} = 1$), and the numerator is the probability that a randomly selected person is assigned to the treatment group ($\mathit{Assign} = 1$) and then gets cured ($\mathit{Cure} = 1$). This unknown conditional probability can be easily estimated—by the observed proportion of the cured individuals in the treatment group. The other three conditional probabilities can be similarly estimated by observed proportions. This procedure for estimating the conditional probabilities on the right-hand side of the equation, and thus estimating the LATE on the left-hand side, is known as *instrumental variable estimation* (with the variable Assign serving as the so-called instrument).

This result marks an important achievement. Recall that the LATE is defined in counterfactual terms, using the contents of cards that cannot all be flipped to reveal their faces at the same time—a single person cannot simultaneously take the treatment and not take it. Fortunately, to estimate the LATE, it suffices to observe some proportions in the treatment and control groups and estimate the counterfactual-free, conditional probabilities on the right-hand side of the equation in Theorem 1. It is amazing that an interesting quantity defined in counterfactual terms (the LATE on the left) can be *identified* with a quantity that depends solely on counterfactual-free probabilities (on the right), which are easy to estimate. Thus, this theorem is also known as an *identification* result. Many important theorems in statistics and econometrics for causal inference are identification results.

For a rigorous statement of Theorem 1, see Appendix A, which seeks to improve upon standard presentations. To be sure, there is a particularly lucid and frequently

cited presentation in the statistics article by Angrist, Imbens, & Rubin (1996, Proposition 1), but those authors list only four assumptions, omitting an explicit statement of CEM. In Appendix A, I identify eight assumptions in total, including CEM, of course.

This concludes the first task of this paper: a crash course on the Rubin causal model and the identification result for the LATE.

3 Playing Good Cop

The preceding discussion can inspire a new argument for Conditional Excluded Middle. Let me flesh it out, playing the role of the good cop—for now.

3.1 A New Argument for CEM

Why might it be interesting to have a new argument supporting CEM? The reason is that there is a highly influential argument against CEM (Lewis 1973). Let me briefly review it. Consider the following pair of sentences:

(A) If i took the treatment, i would be cured.
(B) If i took the treatment, i would not be cured.

CEM requires that the disjunction $(A) \vee (B)$ is true in every possible world. To find a counterexample, consider an indeterministic world in which the following holds:

(C) If i took the treatment, i would have a nontrivial probability p of being cured and a probability $1 - p$ of being not cured, where nontriviality means that p lies strictly between 0 and 1.

Then argue as follows that the truth of (C) implies the falsity of both (A) and (B):

Indeterminist Argument Against CEM

1. Assume that (C) is true.

2. By 1, if the individual i took the treatment, i would have a more-than-zero probability of being not cured.

3. So, if i took the treatment, i could be not cured. (This follows from 2, by the inference from 'would have a more-than-zero probability to be' to 'could be'.)

4. Now, suppose for *reductio* that (A) is true: if i took the treatment, i would be cured.

5. Then, by 3 and 4, we have: if i took the treatment, i would be cured *and* could be not cured—absurd.

6. So, by the *reductio* argument from 4 to 5, it follows that (A) is false.

7. By symmetry, (B) is false, too; thus (A) and (B) are both false.

In a nutshell, nontrivial counterfactual probabilitiy refutes CEM—or so Lewis (1973) concludes. Hájek (manuscript) further argues that such counterexamples to CEM are pervasive in the actual world we live in.

The above is just round one of the debate. The next round features responses from defenders of CEM, such as Stalnaker (1981). This debate has unfolded across philosophy of language (Williams 2010), metaphysics (Emery 2017), and traditional epistemology (Boylan 2024).[2] I submit that philosophy of science is also an area where we can explore a new argument for CEM:

Indispensability Argument For CEM

> CEM is assumed in our best theory of causal inference in health and social sciences, whose application to instrumental variable estimation underpinned the 2021 Nobel Prize in Economics. Despite the influential challenge raised by statistician Dawid (2000) more than twenty years ago in the scientific community —a challenge very similar to Lewis's (1973) worry from nontrivial counterfactual probability—CEM has persisted as a core assumption of this theory to this day. Thus, CEM seems indispensable. Given that we should believe in our best theory of causal inference in health and social sciences, and that CEM is an indispensable part of it, it seems that we have no choice but to believe in CEM—for fear of *intellectual dishonesty*, in Putnam's (1971) terms.

As just mentioned, the indispensability of CEM is already supported by its persistence in the fact of the challenge in the scientific community. This indispensability can be further reinforced by examining the role of CEM in the Rubin causal model, to which I turn now.

[2]For reviews of this debate, see Loewenstein (2021) and Mandelkern (2022, sec. 17.3.4).

3.2 What's the Role of CEM, Exactly?

The answer lies in some lemmas proved using the Rubin causal model. At this point, we must examine some formal aspects of the model.

Let $Take_i = 1$ express the proposition that the individual i takes the treatment. Similarly, $Cure_i = 1$ expresses that i is cured, and $Assign_i = 1$ expresses that i is assigned to the treatment group (rather than the control group). To this notation, we can add superscripts to express counterfactuals, such as the following:

- $Cure_i^{Take_i=1} = 1$ means that i would be cured if i took the treatment.
- $Take_i^{Assign_i=0} = 0$ means that the individual i would not take the treatment if i were assigned to the control group.

The Rubin causal model makes some logical assumptions:

> **Assumption (Centering).** The antecedent of a counterfactual is redundant if it happens to be true; in symbols:
>
> $$X_i = x \;\Rightarrow\; \left(Y_i^{X_i=x} = y \;\Leftrightarrow\; Y_i = y\right).$$

There is another logical assumption, being the focus of this paper:

> **Assumption (Conditional Excluded Middle, or CEM).** If Y_i is a binary variable, so is the counterfactual variable $Y_i^{X_i=x}$.

Notably, presentations in the scientific literature almost always mention only in passing that $Y_i^{X_i=x}$ is a binary variable, without recognizing that this is a substantive assumption. The substance can be appreciated only by going from the formalism back to the intended interpretation: To say that $Cure_i^{Take_i=1}$ is binary is to say that either $Cure_i^{Take_i=1} = 1$ or $Cure_i^{Take_i=1} = 0$, which means that either i would be cured under the treatment or i would not be cured under the treatment—an instance of CEM.

The four cards for each individual i correspond to the four counterfactual variables: $Cure_i^{Take_i=1}$, $Cure_i^{Take_i=0}$, $Take^{Assign_i=1}$, and $Take^{Assign_i=0}$, whose values correspond to the faces of the four cards. Thus, the card-based definitions presented above can be formalized as follows. The ITE (individual treatment effect) for an individual i is defined by:

$$\text{ITE}_i \;=_{\text{df}}\; Cure_i^{Take_i=1} - Cure_i^{Take_i=0}.$$

The four subpopulations are defined as follows:

$$\begin{aligned}
\mathit{Complier}(i) &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 0 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 1\,; \\
\mathit{Defier}(i) &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 1 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 0\,; \\
\mathit{Always\text{-}Taker}(i) &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 1 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 1\,; \\
\mathit{Never\text{-}Taker}(i) &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 0 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 0\,.
\end{aligned}$$

Then we have:

> **Lemma A.** Under the assumption of CEM, the four subpopulations just defined—compliers, defiers, always-takers, and never-takers—are mutually exclusive and jointly exhaustive.

The proof is simple: mutual exclusion follows immediately from the definitions; joint exhaustion follows immediately from the definitions and the assumption of CEM. This role of CEM will become important shortly. The local average treatment effect (for the compliers) is defined as:

$$\text{LATE} \;=_{\text{df}}\; \frac{\sum_{i:\,\mathit{Complier}(i)} \text{ITE}_i}{\#\{i : \mathit{Complier}(i)\}}\,,$$

where the denominator denotes the size of the complier subpopulation.

Here is the key point: For the technical reasons explained in Appendix A.4, existing proofs of the identification result for the LATE (Theorem 1) all require estimating an important quantity: *the proportion of compliers in the population.* It is the estimation of this quantity that makes use of Lemma A, which ultimately relies on CEM. Let me explain how this quantity is estimated to reveal the deeper involvement of CEM.

Let's start by considering the proportions of the four subpopulations:

(1) the proportion of compliers in the population;
(2) the proportion of defiers in the population;
(3) the proportion of never-takers in the population;
(4) the proportion of always-takers in the population.

Those four proportions sum to 1 because the corresponding four subpopulations are mutually exclusive and jointly exhaustive, as stated in Lemma A, which assumes CEM. So, to estimate the primary target, (1), it suffices to subtract the estimates of the

remaining three proportions from 1. This is the *first* role played by Lemma A, and hence, by CEM.

Since (2) is equal to zero by the assumption of No Defiers, it remains to estimate (3) and (4).

To estimate (3), consider the following three quantities:

(3) the proportion of never-takers in the population;

(3′) the proportion of never-takers in the control group;

(3″) the proportion of those who end up not taking the treatment in the control group.

Assuming random assignment to the treatment or control group, proportion (3) can be estimated by proportion (3′), provided that we can obtain an accurate value for the latter. And we can. The idea is to exploit this lemma:

> **Lemma B.** Under the assumptions of CEM, Centering, and No Defiers, it follows that, within the treatment group, the never-takers are exactly those who end up not taking the treatment. Or in symbols, $Assign_i = 1$ implies this equivalence:
>
> $$Never\text{-}Taker(i) \;\Leftrightarrow\; Take_i = 0\,.$$

Thanks to this lemma, proportion (3′) is equal to proportion (3″), which can be easily observed: simply count the number of individuals not taking the treatment in the treatment group and divide it by the size of the treatment group. To recap: CEM is assumed in Lemma B, which enables us to use proportion (3″) as an accurate value of proportion (3′), which, by randomization, can then be used as an estimate of proportion (3). Very ingenious indeed!

The idea behind the proof of Lemma B is also clever, drawing on Lemma A. This is a *second* role played by Lemma A, and hence, by CEM. Let me present the proof in plain language.

Proof of Lemma B. To prove the "⇒" direction, consider any individual i being a never-taker in the treatment group. Then, by the assumption of Centering, i does not take the treatment. Now, to prove the "⇐" direction, consider any individual i in the treatment group who ends up not taking the treatment. By Lemma A, which relies on CEM, this person i must be one of the following: an always-taker, never-taker, defier, or complier—notably, this is the only place where CEM is employed in this proof. Of those four possibilities, three can be eliminated. Specifically, we can eliminate the

possibility that i is a defier, by the assumption of No Defiers. We can also eliminate the possibility that i is an always-taker or complier; for the always-takers and complier in the treatment group end up taking the treatment by the assumption of Centering, but i does *not* take the treatment. Thus, the only remaining possibility is that i is a never-taker, as desired. Q.E.D.

Now that we know how to estimate proportion (3), the same trick can be used to estimate (4), the proportion of always-takers in the population, by counting the actual takers in the control group, and by applying a similar lemma with a similar proof.[3] And recall that proportion (2) equals zero. Once estimates of proportions (2), (3), and (4) are obtained as explained above, subtracting them from 1 yields an estimate of the primary target, (1), the proportion of compliers.

To wrap up: While the assumption of CEM is often obscured in the formalism of the Rubin causal model, I hope the above reconstruction illuminates the deeply involved roles that CEM plays in both the model and its applications to causal inference. No wonder CEM has remained a core assumption for more than twenty years even after the influential challenge posed by the statistician Dawid in 2000 within the scientific community. This strongly suggests that CEM is indispensable to our best theory of causal inference in health and social sciences.

I have thus completed my second task: presenting a new argument that proponents of CEM can explore and utilize—an indispensability argument drawn from the 2021 Nobel Prize in Economics. To further the dialectic, it is now time for me to switch sides and assist opponents of CEM.

4 Playing Bad Cop

In my role as the bad cop, I argue that the above theory of causal inference can be reformulated to dispense with CEM.[4] To this end, expansion packs to the base game are needed.

[3] This lemma, **Lemma B'**, states that, under the assumptions of CEM, Centering, and No Defiers, we have that, within the control group, the always-takers are exactly those who end up taking the treatment; in symbols, $Assign_i = 0$ implies this equivalence: $Always\text{-}Taker(i) \Leftrightarrow Take_i = 1$.

[4] This approach is akin, at least in spirit, to what Field (1980/2016) did to challenge the Quine-Putnam indispensability argument for mathematical realism, by reformulating Newtonian mechanics without referring to real numbers.

4.1 Expansion Pack: Going Stochastic

In the base game, everyone is given only a *single* card printed with '*if Take* = 1', whose face determines whether that person would, or would not, be cured under the treatment. But now imagine that you are given not just one card printed with '*if Take* = 1', but a *deck* of such cards, where 80% are printed with '*Cure* = 1' on their faces, and the remaining 20% with '*Cure* = 0'. Let this deck be thoroughly shuffled, with all faces down initially. What if you took the treatment? Nature would then randomly draw a card from this deck and flip it to reveal your medical result. Consequently, you would have an 80% probability of being cured.[5] So, you could be cured and could be not cured, and hence, it is neither true that you would be cured nor that you would not be cured. CEM is thereby rendered invalid—or so the Lewisians contend.

Let's generalize. In the base game, every individual is given four cards, each answering one of the following what-if questions:

- What if one took the treatment?
- What if one didn't take the treatment?
- What if one were assigned to the treatment group?
- What if one were assigned to the control group?

Now, let each individual's four cards be replaced by four decks, which provide answers in the following form: 'If individual i were ..., then i would have a probability p of being ...'. Such a p is a counterfactual probability—a probability under a counterfactual condition.

So, we now have a stochastic version of the Rubin causal model: single cards are replaced by decks of cards—that is, deterministic outcomes are replaced by counterfactual probabilities. These counterfactual probabilities can then be used to redefine several concepts in the original Rubin causal model.

Start with the ITE (individual treatment effect). Each individual i still has an ITE, but it is now redefined as the difference between two counterfactual probabilities, or equivalently, two proportions in decks of cards:

$$\begin{aligned} \text{ITE}_i \;=_{\text{df}}\; & \text{(i) the proportion of '}\textit{Cure} = 1\text{' cards in } i\text{'s deck for '}\textit{if Take} = 1\text{'} \\ & - \text{(ii) the proportion of '}\textit{Cure} = 1\text{' cards in } i\text{'s deck for '}\textit{if Take} = 0\text{'}. \end{aligned}$$

[5] If randomly drawing a card from a deck does not sound chancy enough, replace it with measuring an observable in a quantum-mechanical system.

In the limiting case where each deck contains only one card, the newly defined ITE reduces to the original ITE.

Subpopulations are redfined, too. Every individual i now has a *degree of compliance* DC_i, defined by how one's counterfactual probability of taking the treatment would change if one switched from the control group to the treatment group:

$$\begin{aligned}\mathrm{DC}_i \;=_{\mathrm{df}}\;& (a) \text{ the proportion of '}\mathit{Take} = 1\text{' cards in } i\text{'s deck for '}\mathit{if\ Assign} = 1\text{'}\\ & - (b) \text{ the proportion of '}\mathit{Take} = 1\text{' cards in } i\text{'s deck for '}\mathit{if\ Assign} = 0\text{'}.\end{aligned}$$

The difference between term (a) and term (b) can be positive, zero, or negative, corresponding to three subpopulations:

- If $\mathrm{DC}_i > 0$, one is called a *complier* (in the general sense).
- If $\mathrm{DC}_i < 0$, one is called a *defier* (in the general sense).
- If $\mathrm{DC}_i = 0$, one is called an *indifferent-taker*, with two special cases: an *always-taker*, who has $(a) = (b) = 100\%$, and a *never-taker*, who has $(a) = (b) = 0\%$.

As to the target of estimation, LATE, it is replaced by a more general concept: a weighted average of the individual treatment effects, where each individual's weight w_i is proportional to their degree of compliance DC_i. This new concept is called the *d*egree-of-compliance-weighted *a*verage *t*reatment *e*ffect, or DATE for short. In symbols:

$$\begin{aligned}\mathrm{DATE} \;&=_{\mathrm{df}}\; \sum_{i\text{: being a complier}} w_i\,\mathrm{ITE}_i\,,\\ w_i \;&=_{\mathrm{df}}\; \frac{\mathrm{DC}_i}{\sum_{j\text{: being a complier}} \mathrm{DC}_j}\,,\end{aligned}$$

where the denominator in the definition of weights w_i is a normalizing factor introduced to ensure that the weights sum to 1.

The present setting is quite general, encompassing the original card game as a limiting case, where every deck contains only one card. In this special case, all compliers are equally compliant, with a maximal degree of compliance (100% minus 0%), which reduces the DATE to the LATE.

4.2 The Final Expansion Pack: Adding a Causal Bayes Net

The next step is to state a key assumption in instrumental variable estimation, which, when expressed in plain language, asserts the following:

> **Assumption (Instrumentality, Informal Version).** The assignment mechanism (to the treatment/control group) causally influences the medical outcome only through whether an individual takes the treatment. Moreover, there is no common cause shared by the assignment mechanism and the medical outcome.

When this assumption holds, the variable *Assign* is called an *instrument*. This informal statement is often found in textbooks (Hernán & Robins 2023, sec. 16.1), but interestingly, the standard formalization of this assumption in the Rubin causal model appears quite different, as you can see from the statement of Assumption 2 in Appendix A (see also Hernán & Robins 2023, technical point 16.1).

I propose a more straightforward formalization of this assumption, using the causal structure depicted in Figure 1.

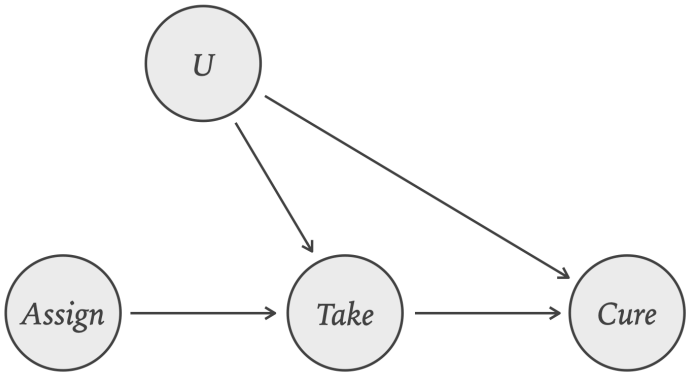

Figure 1: The causal structure that captures the Instrumentality assumption

This causal structure is an exact representation of the Instrumentality assumption: every path from the *Assign* variable to the *Cure* variable passes through the *Take* variable, and there is no common cause shared by *Assign* and *Cure*. The confounding variable, U, is set to be as fine-grained as possible to avoid missing any confounding factors: its possible values are the individuals in the population. This suffices to encompass all the social, economic, and health conditions of each individual.

Next, let's turn this causal graph into a *causal Bayes net*.[6] This is done by specifying some probabilities: the probability distribution of each exogenous variable (i.e., U and *Assign*), and the conditional probability distribution of each effect variable given its direct cause variables, as shown in Figure 2.

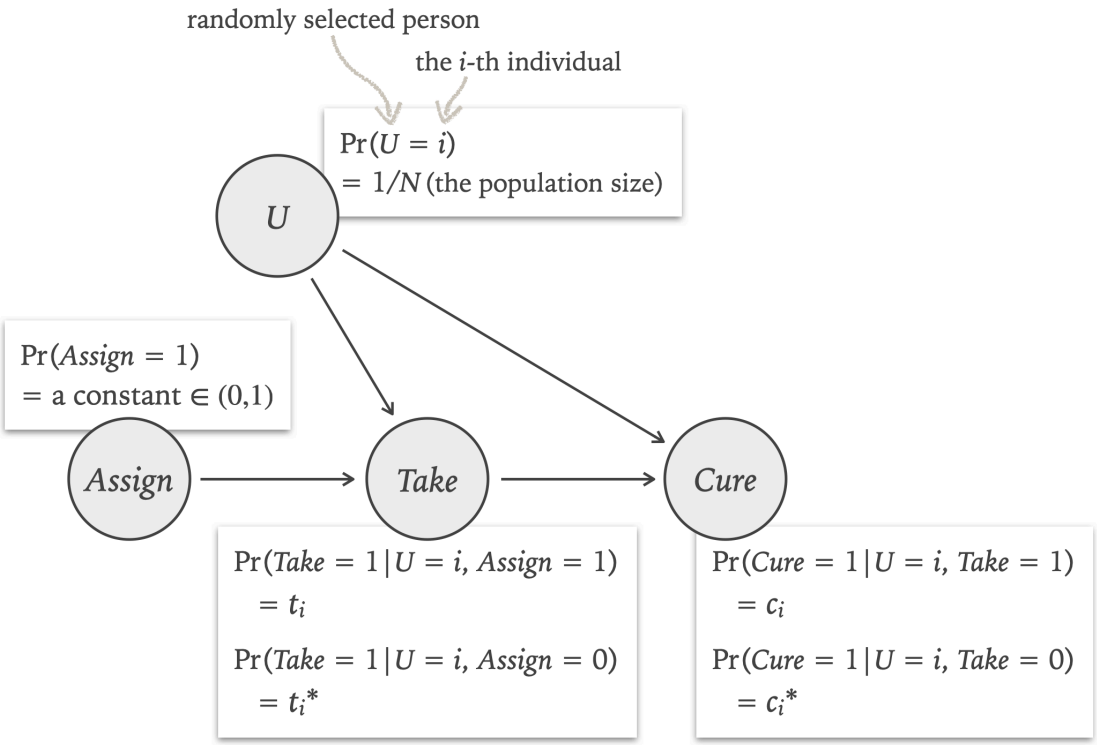

Figure 2: The causal Bayes net assumed in Theorem 2

Those probabilities are defined as follows. First, everyone in the population has an equal probability of being selected, so $\Pr(U = i) = 1/N$, where i is the i-th individual and N is the population size. Once a person i is selected, a coin is flipped to decide whether to assign that person to the treatment or control group, with $\Pr(Assign = 1) = 1/2$, or more generally, $\Pr(Assign = 1)$ being a constant, independent of the individual selected. Finally, the conditional probabilities of effects given direct causes are identified with the appropriate counterfactual probabilities c_i, c_i^*, t_i, and t_i^* (as shown in the Figure 2), whose values are taken from the stochastic version of the Rubin causal model, or equivalently, the stochastic expansion pack to the base

[6]The word 'Bayes' can be misleading. Despite the established name in the literature, there is nothing inherently Bayesian in causal Bayes nets, also known as causal Bayesian networks. The probabilities in such networks are most naturally interpreted as physical objective probabilities, measuring the propensities or tendencies of causal influences, rather than degrees of belief.

game:

$$
\begin{aligned}
t_i &=_{\text{df}} \text{ the proportion of '}\mathit{Take} = 1\text{' cards in } i\text{'s deck for for '}\mathit{if}\ \mathit{Assign} = 1\text{'};\\
t_i^* &=_{\text{df}} \text{ the proportion of '}\mathit{Take} = 1\text{' cards in } i\text{'s deck for for '}\mathit{if}\ \mathit{Assign} = 0\text{'};\\
c_i &=_{\text{df}} \text{ the proportion of '}\mathit{Cure} = 1\text{' cards in } i\text{'s deck for for '}\mathit{if}\ \mathit{Take} = 1\text{'};\\
c_i^* &=_{\text{df}} \text{ the proportion of '}\mathit{Cure} = 1\text{' cards in } i\text{'s deck for for '}\mathit{if}\ \mathit{Take} = 0\text{'}.
\end{aligned}
$$

The main idea can be summarized as follows:

> **Proposal of a New Causal Modeling.** While the original Rubin causal model allows only deterministic outcomes for an individual, it is updated with an expansion pack—replacing single cards with decks—to allow stochastic outcomes with nontrivial counterfactual probabilities. These probabilities are then incorporated into an appropriate causal Bayes net.

This is a combination of two frameworks for causal modeling: the Rubin causal model, more familiar to health and social scientists, and the causal Bayes net, more familiar to philosophers and computer scientists. You will soon see that these two causal models are stronger together, at least for the purpose of pursuing freedom from CEM.

4.3 Dispensing with CEM

Finally, we arrive at a new theorem—a stochastic counterpart to the identification result for the LATE:

> **Theorem 2 (Identification of the DATE).** *Suppose that the following assumptions hold:*
>
> - (Random Selection) *Individuals are randomly selected from the population with equal probabilities.*
> - (Random Assignment) *The selected people are randomly assigned to the treatment or control group with a constant bias strictly between* 0 *and* 1 (*e.g., by flipping a fair coin*).
> - (Instrumentality*) *The true causal model is the causal Bayes net depicted in Figure 2.*
> - (Existence of Compliers*) *There are compliers in the population, in the sense that someone's degree of compliance is positive.*

- (No Defiers*) *There are no defiers in the population, in the sense that no one's degree of compliance is negative.*

Then the DATE *can be expressed solely in terms of probabilities over the observable variables—Assign, Take, and Cure—without counterfactuals. Specifically:*

$$\text{DATE} \;=\; \frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)}\,.$$

See Appendix B for a proof. The first two assumptions are actually redundant, as they are already encapsulated in the causal Bayes net posited in the third assumption; but they are stated here to highlight the role of randomization. The last three assumptions are labeled with asterisks to distinguish them from their counterparts in the original Rubin causal model, as stated in Appendix A.

This new theorem has a notable feature: the right-hand side of the equation for the DATE in the new theorem is *identical* to that for the LATE in the classic result. Both are expressed as the same combination of conditional probabilities: $\frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) - \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) - \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)}$. This feature is crucial. Scientists can continue using the same procedure of instrumental variable estimation—estimating the left-hand side by estimating the exact same conditional probabilities on the right-hand side, based on the exact same proportions observed in the treatment and control groups. However, thanks to this new theorem, the old estimation procedure no longer assumes CEM and can be *reinterpreted* as estimating the new left-hand side: the newly defined causal effect DATE, of which the LATE is merely a limiting case in a deterministic world (at least for Lewisans).

This reinterpretation undermines the indispensability argument. Medical and social scientists have practiced instrumental variable estimation for decades, with the stated goal of estimating the LATE under the assumption of CEM. Yet this well-established practice can now be reinterpreted as actually estimating the DATE all along—without assuming CEM. So, the successes of the original theory for causal effect estimation are preserved in the new theory, which dispenses with CEM. The indispensability argument is thus defused.

At this point, proponents of CEM might reply that even if they are compelled to adopt the new theory of causal inference, this would not to stop them from holding onto CEM. Indeed, the assumptions of the new theory only involve counterfactual probabil-

ities and do not explicitly refer to the logic of counterfactuals. And Stalnaker (1981) already argued that one can coherently embrace nontrivial counterfactual probabilities and insist on CEM at the same time. The idea is based on a semantic technique known as *supervaluation*, used to resist Lewis's (1973) argument that nontrivial counterfactual probability refutes CEM.

Setting aside the details of supervaluation, it suffices to note that I, as the bad cop at this moment, can concede the points that Stalnakerians made in the previous paragraph. Even so, my main point remains: thanks to the new theory of causal inference, CEM is no longer *indispensable*, even if it might still be *optional*. This is sufficient to undermine the indispensability argument—the mere optionality of an option is too weak to entail that we should take that option.

This concludes my role as the bad cop, whose goal is to defuse the indispensability argument.

5 Closing

The Rubin causal model, with its underlying logic of counterfactuals and its capacity to facilitate causal inference, merits closer examination by philosophers. To this end, the preceding discussion provided a card-game tutorial to introduce the model and developed a dialectic connecting it to familiar philosophical issues. The focus was on the ongoing debate surrounding a logical principle: Conditional Excluded Middle (CEM). I developed both sides of the debate in turn. First, I demonstrated how the Rubin causal model could be used to construct a new argument for CEM—a Quine-Putnam-style indispensability argument. Then, I switched sides and challenged this argument by employing a causal Bayes net that renders CEM dispensable.

Where does my heart go to, the Stalnakerian side for CEM, or the Lewisian side against it? I find myself leaning slightly toward the latter, but that is secondary for now. The real takeaway is how the dialectic between the good cop and the bad cop underscores the intriguing potential of the Rubin causal model for philosophers. Indeed, I see opportunities for both sides of the debate.

For proponents of CEM, the next step might be to explore whether the use of causal Bayes nets is most sensible with or without CEM. Opponents of CEM, on the other hand, should further explore the potential of causal Bayes nets as an enhancement of the original Rubin causal model, extending beyond the application to instrumental

variable estimation as discussed above. After all, the original Rubin causal model has several significant applications, such as the method of *difference in differences*; see Hernán & Robins (2023, technical points 7.3 and 21.13).

The dialectic above also suggests an interesting case for the revisability of logic. If health and social scientists can be persuaded to abandon CEM, possibly following the new theory of causal inference developed above, it would be an example of how empirical inquiry can drive revisions in deductive logic—precisely the kind of case Quine (1951) envisioned. This would underscore the possibility of revising logic based not only on empirical inquiries but specifically on those addressing immediate practical concerns, such as in the health and social sciences—a much more relatable example than Putnam's (1968) proposal to shift from classical to quantum logic.

So much for deductive logic, but there is also something here for theorists of induction. When scientists justify inductive methods, they rely heavily on their contexts of inquiry, including background assumptions. Past discussions have mostly focused on background assumptions that are physical (Longino 1979, Christensen 1997), methodological, or ethical (Reiss 2020), rather than logical. But do scientists have to assume a logical principle like CEM to justify certain causal inferences? As we have seen, the search for an answer is far from trivial. Thus, background assumptions about deductive logic warrant greater attention from theorists of induction.

There is also something for those more interested in scientific modeling rather than inference, whether deductive or inductive. Consider the interplay between three approaches to causal modeling:

(1) Rubin causal models (Rubin 1974),
(2) structural equation models (Pearl 2009),
(3) causal Bayes nets (Spirtes et al. 2000).

Pearl (2009) famously argues that the first two approaches—Rubin causal models and structural equation models—are essentially equivalent and can produce everything we can do with the third approach: causal Bayes nets. However, the new theorem suggests a different picture: in at least one important application (instrumental variable estimation), causal Bayes nets appear to generalize Rubin causal models with an extended result. This prompts a reconsideration of these questions: Which approach to causal modeling is more general? Which are equivalent, and in what sense? These questions would make for an interesting case study on an important topic: intertheory relations,

a subject whose case studies have thus far been largely drawn from natural sciences.[7] I submit that more attention be directed to the relations among causal models in health and social sciences. While initial steps have been taken by Markus (2021) and Weinberger (2023), their work does not consider causal Bayes nets. Much more remains to be explored.

The Rubin causal model clearly offers a rich landscape for further exploration by philosophers.

Acknowledgements

I am indebted to the participants of the Workshop on the Philosophy, Psychology, and Computer Science of Causation held in Kyoto (June 24-26, 2023, Kyoto, Japan), the Conference on Causality in Epidemiology (May 2-4, 2024, Linz, Austria), and the Causation session at the 2024 Philosophy of Science Association Biennial Meeting (November 14-17, 2024, New Orleans, LA, USA). I am especially grateful to Christopher Hitchcock, Peng Ding, Jiji Zhang, Frederick Eberhardt, Jun Otsuka, Xiao-Li Meng, Konstantin Genin, Conor Mayo-Wilson, Tom Wysocki, and Jennifer Jhun for stimulating questions and discussions.

Appendices

A The Formalism of the Rubin Causal Model

The Rubin causal model builds on a simple idea: ordinary variables are extended to variables under counterfactual conditions, also known as *potential outcomes*.

A.1 The Potential Outcome Notation

Recall that $\mathit{Take}_i = 1$ expresses the proposition that individual i takes the treatment. Similarly, $\mathit{Cure}_i = 1$ says that i gets cured, and $\mathit{Assign}_i = 1$ says that i is assigned to the treatment group (rather than the control group). Given a variable X_i, we can use X_i^{C} to denote a *potential outcome*, which represents the value of X that individual i would have under the counterfactual condition C. For example:

[7] For a review of this subject, see Palacios (2024).

- $Cure_i^{Take_i=1} = 1$ means that individual i would be cured if i took the treatment.
- $Cure_i^{Take_i=1, Take_j=0} = 0$ means that individual i would not be cured if i took the treatment but another individual, j, did not.

There are eight assumptions for instrumental variable estimation, beginning with those that are inherently tied to the card design introduced in the tutorial.

A.2 Assumptions Built into the Card Design

Here is a substantive, empirical assumption:

> **Assumption 1 (Stable Unit Treatment Value, or SUTVA).** The values of the variables of each individual (or unit) are determined independently of the values of the variables of any other individuals. That is, for any variable X_i and any conditions $\mathrm{C}_1, \ldots, \mathrm{C}_n$ concerning individuals from 1 to n, we have $X_i^{\mathrm{C}_1,\ldots,\mathrm{C}_n} = X_i^{\mathrm{C}_i}$, which omits any references to individuals other than i in the counterfactual condition.

This assumption might be violated in some cases, such as when dealing with a contagious disease in a densely populated community. However, when it is plausible to make this assumption, the analysis becomes simpler: to determine whether i would be cured under various conditions, it suffices to consider potential outcomes of the form $Cure_i^{Assign_i=a, Take_i=t}$.

The next assumption enables further simplification by dropping additional terms from the counterfactual conditions:

> **Assumption 2 (Instrumentality).** For each individual i, $Assign_i$ is an instrumental variable in the following sense: the value of $Cure_i$ is determined once the value of $Take_i$ is determined, independently of the value of $Assign_i$. That is, $Cure_i^{Assign_i=a, Take_i=t} = Cure_i^{Take_i=t}$, which omits the assignment $Assign_i = a$ in the counterfactual condition.

Thanks to the above two assumptions, now we only need to consider just four potential outcomes for each individual i: $Cure_i^{Take_i=1}$, $Cure_i^{Take_i=0}$, $Take_i^{Assign_i=1}$, $Take_i^{Assign_i=0}$. Those four variables correspond to the four cards that i holds in the game presented in the tutorial (Section 2). The faces of the four cards are printed with the values of those four potential outcomes, respectively. By choosing to model each individual with

such a simple set of four cards, we are already committed to the substantive *empirical* assumptions of SUTVA and Instrumentality.

The design of four cards (as opposed to the four-deck design in my expansion pack) also carries substantive assumptions, but these are *logical* assumptions this time. The first is:

> **Assumption 3 (Centering/Consistency).** It must be that
>
> $$X_i = x \;\Rightarrow\; \left(Y_i^{X_i=x} = y \;\Leftrightarrow\; Y_i = y\right);$$
>
> that is, an antecedent $X_i = x$ in a counterfactual is redundant if it turns out to be true.

While 'Centering' is the standard name for this logical principle in philosophy, the scientific literature uses 'Consistency' instead. There is a second logical assumption, which is the focus of this paper:

> **Assumption 4 (Conditional Excluded Middle, or CEM).** Suppose that Y_i is a binary variable. Then the counterfactual variable $Y_i^{X_i=x}$ is also a binary variable; in other words:
>
> $$Y_i^{X_i=x} = 1 \;\vee\; Y_i^{X_i=x} = 0\,.$$

This disjunctive schema represents CEM as it is understood in philosophy.

Under the assumption of CEM, the four subpopulations defined below are mutually exclusive and jointly exhaustive (as stated in Lemma A in Section 3.2):

$$\begin{aligned}
\mathit{Complier}(i) &\;\Leftrightarrow_{\text{df}}\; \mathit{Take}_i^{\mathit{Assign}_i=0} = 0 \;\wedge\; \mathit{Take}_i^{\mathit{Assign}_i=1} = 1\,;\\
\mathit{Defier}(i) &\;\Leftrightarrow_{\text{df}}\; \mathit{Take}_i^{\mathit{Assign}_i=0} = 1 \;\wedge\; \mathit{Take}_i^{\mathit{Assign}_i=1} = 0\,;\\
\mathit{Always\text{-}Taker}(i) &\;\Leftrightarrow_{\text{df}}\; \mathit{Take}_i^{\mathit{Assign}_i=0} = 1 \;\wedge\; \mathit{Take}_i^{\mathit{Assign}_i=1} = 1\,;\\
\mathit{Never\text{-}Taker}(i) &\;\Leftrightarrow_{\text{df}}\; \mathit{Take}_i^{\mathit{Assign}_i=0} = 0 \;\wedge\; \mathit{Take}_i^{\mathit{Assign}_i=1} = 0\,.
\end{aligned}$$

A.3 An Auxiliary Assumption

Now we can define the target of estimation. The ITE (individual treatment effect) for an individual i is defined by:

$$\text{ITE}_i \;=_{\text{df}}\; \mathit{Cure}_i^{\mathit{Take}_i=1} - \mathit{Cure}_i^{\mathit{Take}_i=0}\,.$$

The local average treatment effect (LATE) for the compliers is defined as:

$$\text{LATE} \;=_{\text{df}}\; \frac{\sum_{i:\,\mathit{Complier}(i)} \text{ITE}_i}{\#\{i : \mathit{Complier}(i)\}}\,.$$

To make this target of estimation well-defined, the denominator must be assumed to be nonzero:

> **Assumption 5 (Existence of Compliers)**
> $\mathit{Complier}(i)$ for some individual i.

None of the above assumptions involves probabilities, which will soon make their appearance.

A.4 Probabilistic Assumptions

The design of the cards itself is non-probabilistic. In the Rubin causal model, probabilities arise from how individuals are drawn from the population and assigned to different groups. For simplicity, let the subscript-free notation $\Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1\big)$ denote the probability of drawing an individual from the population who would be cured without taking the treatment. If everyone has an equal probability $1/N$ of being selected, where N is the population size, then $\Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1\big)$ is identical to the proportion of those who would be cured without taking the treatment. This exploits a convenient ambiguity of Pr between *pr*obability and *pr*oportion. Now I cam state the first probabilistic assumption:

> **Assumption 6 (Random Selection).** Everyone in the population has an equal probability of being selected. So the actual frequency distribution of the four potential outcomes in the population is the same as the probability distribution of those variables. In other words, of the individuals with the following counterfactual properties:

- $Take^{Assign=0} = a$,
- $Take^{Assign=1} = b$,
- $Cure^{Take=0} = c$,
- $Cure^{Take=1} = d$

their proportion in the population is equal to:

$$\Pr\big(Take^{Assign=0} = a,\ Take^{Assign=1} = b,\ Cure^{Take=0} = c,\ Cure^{Take=1} = d\big)\,.$$

Thanks to the above assumption, we can derive a probabilistic formula to express the LATE—a formula often treated as a definition in textbooks for convenience, but which is actually a lemma (Imbens & Rubin 2015):

Lemma C. Under the assumption of Random Selection, we have:

$$\text{LATE} \;=\; \Pr\big(Cure^{Take=1} = 1 \,|\, Complier\big) \;-\; \Pr\big(Cure^{Take=0} = 1 \,|\, Complier\big)\,.$$

Proof of Lemma C. Calculate the LATE as follows, where Existence of Compliers is assumed throughout to make all denominators nonzero, and the last step applies the assumption of Random Selection.

$$\begin{aligned}
&\text{LATE} \\
&= \frac{1}{\#\{i : Complier(i)\}} \sum_{i:\, Complier(i)} \text{ITE}_i \\
&= \frac{1}{\#\{i : Complier(i)\}} \sum_{i:\, Complier(i)} \big(Cure_i^{Take_i=1} - Cure_i^{Take_i=0}\big) \\
&= \frac{\sum_{i:\, Complier(i)} Cure_i^{Take_i=1}}{\#\{i : Complier(i)\}} \;-\; \frac{\sum_{i:\, Complier(i)} Cure_i^{Take_i=0}}{\#\{i : Complier(i)\}} \\
&= \frac{\#\{i : Cure_i^{Take_i=1} = 1 \wedge Complier(i)\}}{\#\{i : Complier(i)\}} \;-\; \frac{\#\{i : Cure_i^{Take_i=1} = 0 \wedge Complier(i)\}}{\#\{i : Complier(i)\}} \\
&= \Pr\big(Cure^{Take=1} = 1 \,|\, Complier\big) \;-\; \Pr\big(Cure^{Take=0} = 1 \,|\, Complier\big)\,. \quad \text{Q.E.D.}
\end{aligned}$$

If we unpack the conditional probabilities on the right-hand side using the standard definition, there will appear a denominator, $\Pr(Complier)$, which is equal to the

proportion of compliers in the population, assuming Random Selection. This is why the proof of the theorem for the LATE requires considering how this proportion is estimated, which, in turn, relies on the assumption of CEM, as explained in Section 3.2.

In addition to the assumption of Random Selection, there is a second probabilistic assumption:

> **Assumption 7 (Random Assignment).** Any individual, once selected, has a nontrivial probability (say 50%) of being assigned to the treatment/control group, independently of their identity. So, *Assign* is probabilistically independent of the set of all the four potential outcomes in use, $Take^{Assign=0}$, $Take^{Assign=1}$, $Cure^{Take=0}$, and $Cure^{Take=1}$; in symbols:

$$\begin{aligned}
&\Pr\big(Take^{Assign=0} = a,\ Take^{Assign=1} = b,\ Cure^{Take=0} = c,\ Cure^{Take=1} = d\big) \\
&= \Pr\big(Take^{Assign=0} = a,\ Take^{Assign=1} = b,\ Cure^{Take=0} = c,\ Cure^{Take=1} = d \,|\, Assign = 0\big) \\
&= \Pr\big(Take^{Assign=0} = a,\ Take^{Assign=1} = b,\ Cure^{Take=0} = c,\ Cure^{Take=1} = d \,|\, Assign = 1\big)\ .
\end{aligned}$$

A.5 The Last Mile

There is one final assumption:

> **Assumption 8 (No Defiers)** $Defier(i)$ for no individual i.

This is another substantive empirical assumption and is presented last because, in real applications, it is often the one most responsible for delineating the scope of the method of instrumental variable estimation.

Then we have the classic result due to Imbens & Angrist (1994) and Angrist, Imbens, & Rubin (1996):

> **Theorem 1 (Formal Version).** Under the assumptions 1-8 as stated above,
>
> $$\text{LATE} \ = \ \frac{\Pr(Cure = 1 \,|\, Assign = 1) \ - \ \Pr(Cure = 1 \,|\, Assign = 0)}{\Pr(Take = 1 \,|\, Assign = 1) \ - \ \Pr(Take = 1 \,|\, Assign = 0)}\ .$$

I believe that this list of assumptions, 1-8, is the most comprehensive one currently available.

B Proof of the Main Result: Theorem 2

Recall that each individual i has an individual treatment effect given by:

$$\text{ITE}_i \;=\; c_i - c_i^* \,,$$

with a degree of compliance given by:

$$\text{DC}_i \;=\; t_i - t_i^* \,.$$

Hence the DATE can be expressed as follows:

$$\begin{aligned}
\text{DATE} \;&=\; \sum_{i\text{: being a complier}} \underbrace{\left(\frac{\text{DC}_i}{\sum_{j\text{: being a complier}} \text{DC}_j} \right)}_{=\text{ the weight of } i} \text{ITE}_i \\
&=\; \sum_i \left(\frac{\text{DC}_i}{\sum_j \text{DC}_j} \right) \text{ITE}_i \\
&=\; \sum_i \left(\frac{t_i - t_i^*}{\sum_j \left(t_j - t_j^* \right)} \right) \left(c_i - c_i^* \right) .
\end{aligned}$$

The first line is just the definition of the DATE, which is well-defined (with a nonzero denominator) by the assumption of Existence of Compliers*. In the second line, i and j are no longer restricted to compliers but range over all individuals in the population; this is justified by the assumption of No Defiers* and by the fact that indifference-takers carry zero weights. Now, the goal is to verify this equation:

$$\text{DATE} \;\stackrel{?}{=}\; \frac{\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1) \;-\; \Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0)}{\Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 1) \;-\; \Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 0)} \,.$$

The terms on the right-hand side are to be calculated in turn. I will leverage a defining feature of the causal Bayes net, the *Causal Markov Assumption*, which asserts that every variable is probabilistically independent of its non-descendants (non-effects) given

its parents (direct causes). Start with the first term in the numerator:

$$
\begin{aligned}
&\Pr(\mathit{Cure}=1 \mid \mathit{Assign}=1) \\
&= \sum_{i,j} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=j, U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(\mathit{Take}=j \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i \mid \mathit{Assign}=1) \Big) && \text{by Chain Rule} \\
&= \sum_{i,j} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=j, U=i, \cancel{\mathit{Assign}=1}) \\
&\qquad \times \Pr(\mathit{Take}=j \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i \mid \cancel{\mathit{Assign}=1}) \Big) && \text{by Causal Markov} \\
&= \sum_{i} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=1, U=i) \\
&\qquad \times \Pr(\mathit{Take}=1 \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i) \Big) \\
&\quad + \sum_{i} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=0, U=i) \\
&\qquad \times \Pr(\mathit{Take}=0 \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i) \Big) \\
&= \sum_{i} \Big(c_i \, t_i \, \tfrac{1}{N} \Big) \; + \; \sum_{i} \Big(c_i^* \, (1-t_i) \, \tfrac{1}{N} \Big) \\
&= \tfrac{1}{N} \sum_{i} \Big(c_i \, t_i + c_i^* \, (1-t_i) \Big) .
\end{aligned}
$$

Similarly for the second term in the numerator:

$$
\begin{aligned}
&\Pr(\mathit{Cure}=1 \mid \mathit{Assign}=0) \\
&= \tfrac{1}{N} \sum_{i} \Big(c_i \, t_i^* + c_i^* \, (1-t_i^*) \Big) .
\end{aligned}
$$

Now calculate the first term in the denominator:

$$\begin{aligned}
&\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \\
&= \sum_i \Big(\Pr(\mathit{Take}=1 \,|\, U=i, \mathit{Assign}=1) \cdot \Pr(U=i \,|\, \underbrace{\cancel{\mathit{Assign}=1}}_{\text{by Causal Markov}}) \Big) \\
&= \sum_i t_i \, \tfrac{1}{N} \\
&= \tfrac{1}{N} \sum_i t_i \,.
\end{aligned}$$

Similarly for the second term in the denominator:

$$\begin{aligned}
&\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0) \\
&= \tfrac{1}{N} \sum_i t_i^* \,.
\end{aligned}$$

To finish off, plug the four terms just calculated into the following:

$$\begin{aligned}
&\frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)} \\
&= \frac{\frac{1}{\cancel{N}} \sum_i \Big(c_i \, t_i + c_i^* \, (1 - t_i) \Big) - \frac{1}{\cancel{N}} \sum_i \Big(c_i \, t_i^* + c_i^* \, (1 - t_i^*) \Big)}{\frac{1}{\cancel{N}} \sum_i t_i - \frac{1}{\cancel{N}} \sum_i t_i^*} \\
&= \frac{\sum_i \Big(c_i \, t_i + \cancel{c_i^*} - c_i^* \, t_i - c_i \, t_i^* - \cancel{c_i^*} + c_i^* \, t_i^* \Big)}{\sum_i t_i - \sum_i t_i^*} \\
&= \frac{\sum_i \big(t_i - t_i^* \big) \big(c_i - c_i^* \big)}{\sum_j \big(t_j - t_j^* \big)} \\
&= \sum_i \left(\frac{t_i - t_i^*}{\sum_j \big(t_j - t_j^* \big)} \right) \big(c_i - c_i^* \big) \\
&= \text{DATE} \,.
\end{aligned}$$

Q.E.D.

References

Angrist, J. D. (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80, 313-336.

Boylan, D. (2024) "Counterfactual Skepticism Is (Just) Skepticism", *Philosophy and Phenomenological Research*, 108(1), 259-286.

Christensen, D. (1997) "What Is Relative Confirmation?", *Noûs*, 31(3), 370-384.

Dawid, A. P. (2000) "Causal Inference without Counterfactuals", *Journal of the American Statistical Association*, 95(450): 407-424.

Emery, N. (2017) "The Metaphysical Consequences of Counterfactual Skepticism", *Philosophy and Phenomenological Research*, 94(2), 399-432.

Field, H. (2016) *Science without Numbers*, Oxford University Press.

Hájek, A. (unpublished manuscript) "Most Counterfactuals Are False", URL = <https://philarchive.org/rec/HJEMCA>

Hausman, D. M. (2024) "Philosophy of Economics", Zalta, E. N. & Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), URL = <https://plato.stanford.edu/archives/fall2024/entries/economics/>

Hernán, M. A. & Robins, J. M. (2023) *Causal Inference: What If*, Chapman & Hall/CRC.

Hitchcock, C. (2024) "Causal Models", Zalta, E. N. & Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), URL = <https://plato.stanford.edu/archives/sum2024/entries/causal-models/>

Imbens, G. W., & Angrist, J. (1994) "Identification and Estimation of Local Average Treatment Effects", *Econometrica* 62, 467-476.

Imbens, G. W., & Rubin, D. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Lewis, D. K. (1973) *Counterfactuals*, Blackwell.

Longino, H. E. (1979) "Evidence and Hypothesis: An Analysis of Evidential Relations", *Philosophy of Science*, 46(1), 35-56.

Mandelkern, M. (2022) "Modals and Conditionals", in Altshuler, D. (ed.) *Linguistics Meets Philosophy*, Oxford University Press, pp. 502-533.

Markus, K. A. (2021) "Causal Effects and Counterfactual Conditionals: Contrasting Rubin, Lewis and Pearl", *Economics & Philosophy*, 37(3), 441-461.

Palacios, P. (2024) "Intertheory Relations in Physics", Zalta, E. N. & Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), URL = <https://plato.stanford.edu/archives/spr2024/entries/physics-interrelate/>

Pearl, J. (2009), *Causality*, Cambridge University Press.

Putnam, H. (1968) "Is Logic Empirical?", in Cohen, R. S. & Wartofsky, M. W. (eds.) *Boston Studies in the Philosophy of Science*, Vol. 5, D. Reidel: 216-241.

——— (1971) *Philosophy of Logic*, Routledge.

Quine, W. V. (1948) "On What There Is", *Review of Metaphysics*, 2(5): 21-38.

——— (1951) "Two Dogmas of Empiricism", *Philosophical Review*, 60: 20-43.

Reiss, J. (2020) "What Are the Drivers of Induction? Towards a Material Theory^{+}", *Studies in History and Philosophy of Science Part A*, 83, 8-16.

Rubin, D. B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66: 688-701.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000) *Causation, Prediction, and Search*, MIT Press.

Stalnaker, R. C. (1968) "A Theory of Conditionals", in Harper, W. L., Pearce, G. A., & Stalnaker, R. C. (eds.) *Ifs: Conditionals, Belief, Decision, Chance and Time*, Springer Netherlands: 41-55.

Stalnaker, R. (1981) "A Defense of Conditional Excluded Middle", in: Harper, W. L., Pearce, G. A., & Stalnaker, R. (ends), *Ifs: Conditionals, Belief, Decision, Chance and Time*, D. Reidel Publishing Company, pp. 87-104

Weinberger, N. (2023) "Comparing Rubin and Pearl's Causal Modelling Frameworks: A Commentary on Markus (2021)", *Economics & Philosophy*, 39(3), 485-493.

Williams, J. R. G. (2010) *Defending Conditional Excluded Middle*, Noûs, 44(4), 650-668.