Getting Off the Gold Standard: A Holistic Approach to Causal Inference with Entropic Causal Graphs

Robert Kubinec[1,*]

June 1st, 2022

Abstract

While many classify studies as either descriptive or causal, I argue that causality is a continuous construct, and different inference modes–experimental, observational and mechanistic–can at best provide only partial causal information. To discriminate between the relative value of different inference modes, I employ statistical entropy as a possible yardstick for evaluating research designs as different operations on causal graphs. Rather than dichotomize studies as either causal or descriptive, the concept of entropy instead emphasizes the relative causal knowledge gained from a given research finding. I employ this theory to clarify why and when researchers relied on divergent modes of inference to determine the efficacy of vaccines over the course of the COVID-19 pandemic.[1]

[1] Social Science Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[*] Correspondence: Robert Kubinec <rmk7@nyu.edu>

[1]A reproducible version of this paper with code is available at https://github.com/saudiwin/causality/blob/master/gold_standard_covid.Rmd . I thank David Waldner, Christopher Winship, Michael Poznansky, Kevin Munger, Andrew Gelman and Kosuke Imai for helpful comments on this manuscript.

> You shall not crucify mankind upon a cross of gold.
>
> – William Jennings Bryan, July 8, 1896

This paper inverts the meaning of *gold standard* with respect to causality by considering the origin of the phrase in heated monetary debates of the 19th century. The quotation above by William Jennings Bryan referenced devaluation occurring in the United States as a consequence of pegging the dollar to gold reserves. When the price of gold increased, the money supply constricted and borrowers would have to pay more than they initially agreed to. Bryan and his confederates wanted to switch to the silver standard, of which there was a much more plentiful supply, in order to maximize much-needed specie for the United States' quickly growing economy.

The analogy made in this paper is that employing a gold standard in causality can cause a similar deflation of the evidence available to answer a research question. The underlying problem is the need to make a binary decision about whether a piece of research is considered causal or descriptive. Binary decision problems concerning p-values have been rigorously criticized in recent years for obscuring uncertainty in statistical results (Amrhein, Greenland, and McShane 2019), and I apply the same thinking to our decisions about whether we should consider research causal or descriptive. Re-conceptualizing causality as a latent scale, of which the different research styles are observable indicators, can shift the conversation away from dichotomies and towards an appreciation of the relative amount of causal knowledge obtained from a given piece of research.

Controversially, I will make the claim in this paper that *causal identification* is not in fact necessary for causal learning. This statement is not to disparage existing theorems about identification, nor to challenge the utility of the literature for research. Rather, causal identification is a sufficient condition for causal learning, but it is not, in fact, necessary. It is unlikely that we would be able to answer all causal queries without strict causal identification, but it is not implausible, and has occurred in recent areas of research, as I explain in this article.

I argue instead that there are three main ways that scholars can obtain causal knowledge: the counterfactual mode of inference (which I also link to the closely-related manipulationist mode),

the Humean mode of inference, and the mechanistic mode of inference. Each of these modes can be seen as a kind of *silver* standard, incorporating much of what we mean by the term causal, but failing at the same time to encompass all of what we mean (if we could somehow express it).

I propose that a helpful, possibly unifying, *continuous* criterion for evaluating studies across these different modes is that of reducing entropy. Entropy is a framework widely used in statistics to represent the relative amount of information present in a random variable (Shannon 1948). Entropy is a characteristic of a random variable's distribution; the flatter (more uncertain) the distribution, the more entropy exists. When applied to causal graphs, entropy can help describe how diverse types of research can yield varying amounts of causal knowledge. In this paper I use entropy as a criterion by which to evaluate the relative utility of divergent research designs without having to resort to binary classifications about causal versus exploratory research.

I show that applying entropy metrics to causal graphs returns intuitive results that provide a foundation for considering the value of different research designs over the same causal process. In a given causal graph, it is entirely possible that an observational analysis, which can only show associations, will reduce the entropy of a causal graph as much as, or more than, an experiment that is causally identified. This result closely matches many researchers' intuitions, but has not to my knowledge been previously formalized. As the aim of the social sciences should be to reduce our uncertainty in understanding how the social world operates, we would benefit from applying the word causal more liberally rather than to force research designs into a binary mold of descriptive versus causal inference.

1 The New Intellectual Battlefield: Causality

The credibility revolution of the past fifteen (or so) years, which argues for the application of the potential outcomes framework and RCTs as a way of measuring the credibility of statistical analyses, has produced a sea change in how political scientists, economists and increasingly others

measure research success. The potential outcome theories behind the credibility revolution date back to the 1970s or even earlier (Fisher 1935; Rubin 1974; Holland 1986), but for whatever reason, the practice of formal randomized experiments did not take off in fields besides psychology until the 2000s (Green and Gerber 2003; Morgan and Winship 2007; Levitt and List 2008).

More recently, a second credibility revolution has swept through social scientific disciplines that long employed experiments, particularly psychology. This second revolution has questioned the use of discretized decision rules, i.e. p-values, as a source of inferring causal inference, and shown that many published experiments fail to replicate even if the original experiment had statistically significant results (Open Science Collaboration 2015; Gelman and Loken 2013). This revolution has emphasized pre-registering research questions (Nosek et al. 2018), sharing data so that conclusions can be replicated or reproduced (Goodman, Fanneli, and Ioannidis 2016), and even more radical changes, such as fundamentally altering the standards used to judge statistical inference (Benjamin et al. 2017; Amrhein, Greenland, and McShane 2019). While the first revolution has dramatically elevated the status of experimental research designs, the second has ironically pointed to deep problems in how RCTs have been implemented and evaluated.

As a result, there remain significant pockets of resentment at the success of the potential outcomes framework. The success of experiments appears to endanger the role of observational studies, whether qualitative or quantitative, as these studies can never meet the stringent criteria imposed by their randomized kin (Beck 2006; Gerber, Green, and Kaplan 2014; Gerstein, McMurray, and Holman 2019). Previously popular methods like large-N time-series cross section models have come under criticism for failing to either estimate *average treatment effects* (ATEs) (Samii 2016; G. Imbens 2018; Gibbons, Serrato, and Urbancic 2017); the causal criterion of the potential outcomes framework, or to account for missing variables and over-time dynamics (Plümper and Troeger 2019). To enforce the distinction, journals increasingly require scholars to avoid "causal" language like "impact" or "effect" when using observational methods (Hernán 2018; Thapa et al. 2020; Yu, Li, and Wang 2019).

This tension has boiled over into published debates, including in a remarkably broad and heated discussion in the 2018 August issue of *Social Science and Medicine*. On one side, Deaton and Cartwright (2018) argue that the emphasis on RCTs as a cure-all for causal inference is over-blown because researchers often ignore the known limitations of their samples by reference to randomization. While some support Deaton and Cartwright, including Gelman (2018) and Sampson (2018), others argue that recent research on understanding treatment heterogeneity and the application of experimental results to novel problems mitigate Deaton and Cartwright's concerns (G. Imbens 2018; Ioannidis 2018). This brewing dispute has all the hallmarks of a noteworthy battle of the minds, although it could create yet another methodological minefield that many researchers fear to tread upon. The modal researcher is not so interested in causal inference debates per se but rather having their research devalued by their methodological betters for failing to follow some rule or procedure.

The main problem, I maintain, underlying these disagreements is an acute problem for social scientists: while we all want to obtain causal knowledge, we do not in fact know exactly what causal knowledge is (Spirling and Stewart 2022). Existing research shows that causal thinking is deeply connected to human thought processes (Sloman and Lagnado 2015). Causality involves assigning meaning to events, an endeavor that in fact is part of the definition of rational thought (Brashier and Marsh 2020; Koslowski 1996). Perhaps because it is so foundational to how we process the world, we also have trouble encompassing precisely what we mean when we say a relation is causal versus spurious.

I next briefly examine what I call three silver standards (i.e, observable indicators) of causality: counterfactual/manipulationist, correlational, and mechanistic inference. The intention of this overview is not to be exhaustive, as that would likely require a book-length treatment, but rather to emphasize how each of these causal paradigms captures a part, though not all, of what we mean by the term causality. I depart from existing writing on these subjects by treating these three research paradigms as distinct expressions (or indicators) of causal knowledge rather than assigning them to a hierarchy.

2 Counterfactuals and Manipulations

While its role in the social sciences was traditionally minor, manipulationalist and counterfactual causal inference has become the go-to reference for understanding causal relations. I consider these two paradigms, though conceptually distinct, to be grouped together as they are both fundamentally discussing the same thing. In brief, the counterfactual inference imagines an alternate world in which the causal factor is not present (Rubin 1974; Holland 1986), or what is known as the potential outcomes of a given unit, while the manipulationist account emphasizes human (or possibly divine) efforts to force causal factors to take on certain values (Woodward 2003; Morgan and Winship 2007). These brilliant formulations are what launched the "credibility revolution" (Angrist and Pischke 2010; G. W. Imbens and Rubin 2015) and compels widespread support for RCTs across the social and biomedical sciences.

It is not necessary, of course, to use randomization within a counterfactual or manipulationist theory of inference: it is simply more difficult to know whether an intervention affected the outcome. In the potential outcomes framework, the concept of ignorability determines when an observational design without randomization is able to meet the standards of a randomized design by making missing potential outcomes ignorable (Przeworski 2009; G. W. Imbens and Rubin 2015). By doing so, the potential outcomes framework also influenced statistics in observational research, as I describe later.

Pearl (2000) significantly expanded the definition and possibilities of manipulationist inference by his introduction of network relations in the form of directed a-cyclic graphs (DAGs) to stand for causal relationships. Pearl's main insight is that formulations of causal relations based on conditional probabilities alone cannot capture all of what is meant by causality because variables can be associated with each other without having any direct causal relationship. By providing a very rigorous definition to the notion that correlation and causation are separate phenomena, he came up with a framework that precisely relates manipulationist ideas of inference to statistical methods of estimating relationships between variables. As I show later, Pearl's framework can extend far

beyond manipulationist inference, but because he explicitly defines causality as interventions on causal graphs, I group him in this paradigm.

To summarize Pearl's approach to causality in brief, imagine that there are a set of variables Z, and a variable of interest X, all of which jointly cause an outcome Y. To come up with a directed acyclic graph (DAG), all of the variables that are causal factors must be pointing towards the outcome Y or on some chain to Y, as in Figures 1 and 2.

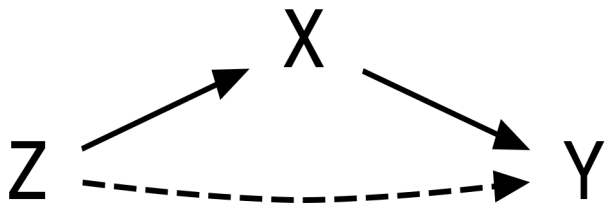

Figure 1: Confounded Example of a Directed Acyclic Graph

Because all the variables are pointing in the same direction, these graphs are acyclic, or each graph can never return to its origin. A causal relation is defined explicitly as fixing one of the causal factors (the *do* operator) so that only the manipulated variable affects the outcome, helpfully removing the confounding association between X and Z in Figure 1 as shown by the box around X in Figure 2.

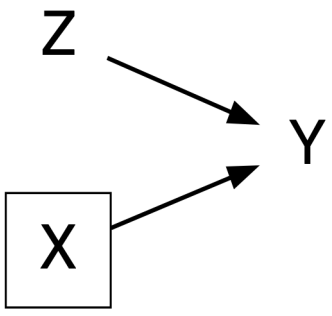

Figure 2: Identified Example of a Directed Acyclic Graph

This manipulation–forcing X to a specific value–removes the influence of Z and the "backdoor path" by which changes in Z cause changes in both X and Y. By visualizing these relationships, and providing an algorithm to convert the graph into observable probabilities, Pearl did a great service to causal inference practitioners. Yet even as powerful as his framework is, it does not capture all of what we mean by causality, at least so far as Pearl defines it. Even though it is possible to use these diagrams for analysis other than direct manipulation, as I show later, it is Pearl who restricts the definition of causality to physical interventions on his networks. While manipulation is certainly a core part of causality, it does not exhaust the subject.

Rather than cast aspersions on RCTs and other forms of counterfactual/manipulationist inference, it is best to think of these methods as providing a very helpful and important component of what is meant by causality. RCTs are one of the best means available for addressing unmeasurable selection attributes, such as situations where wealthier people tend to select into high-income neighborhoods, and vice versa. Under ideal conditions, such as with large samples, low attrition rates, and cleanly measured treatments, RCTs provide a very high amount of causal knowledge.

3 The Relation Between Correlation and Causal Knowledge

Although at one time it was the dominant approach to causal inference in the sciences, the Humean conception of "constant conjunction" has fallen out of fashion. Hume supposed that we could not know why any two events occurred together and to infer any of this knowledge was a fallacy. Rather, all we could know was whether events tended to occur together. This correlational theory of inference was codified by Pearson (Pearl 2018 , 53-91) and has remained a staple of statistical analysis: checking for associations between variables, or looking for risk factors, as in medicine (Boyko and Alderman 1990; Gershman, Guo, and Dahabreh 2018). While today's statistical education emphasizes that correlation does not equal causation, that tendency has not always been as strong among statistical practitioners.

The reason that correlation still matters even if we do not know the direction of causality or the mechanisms behind the correlation is because causation *does* imply association, of which correlation is a specific metric (Altman and Krzywinski 2015). We may not be able to record the association between two variables due to confounders or selection (collider) bias, but if a causal relationship exists, then we can infer that at some level, somewhere, correlation must be happening: if A, then B. Conversely, if we can prove that two variables are perfectly uncorrelated, accounting for our knowledge of the causal process, then we have obtained good–though not perfect–evidence that a causal relationship probably does not exist.

While some are willing to dismiss traditional statistical methods without randomization or at least manipulation of treatments, entire fields of science are based on purely observational analysis, including astronomy, forensic anthropology, paleontology, and so on. We have never manipulated giant bodies of gas to force them to explode, but we are still fairly sure we know what supernova are (Bethe 1990). The reason for this is that as a silver standard, observational analysis does provide causal knowledge if ideal conditions are met, as is true of randomized experiments. In particular, observational methods provide evidence of–though never fully determine–causal relations when we know which variables are relevant to the outcome and in what way, the data provide accurate measures of these variables, and we have as much data as we might want.

Researchers also often use syllogisms to interrogate models (Gelman and Loken 2013) and reason their way through situations where they cannot collect all the data they want (Spirling and Stewart 2022), such as astronomers' disputes over the location and number of planets. If an orbit reflects certain instabilities, then a planet can be hypothesized to exist even if there is no direct evidence for it (Smith 1989).

The denigration of traditional observational models has counter-intuitively led to the rise a class of models that are observational in nature but not defined as so. These methods, including difference-in-difference and regression discontinuity designs, are confusing to define because they make reference to experimental treatments but treatment assignment is never manipulated by the researcher.

Instead, these models' superiority over other observational methods is that they can be expressed in Rubin's potential outcomes notation in a way that establishes ignorability (Lee 2008; Abadie 2005). However, establishing the conditions of ignorability does not itself make those conditions more likely to be met, and practitioners seem to expect that these models will work "out of the box" better than other models.

Without attention paid to the credibility of the ignorability assumptions, these models are essentially derivations of well-known statistical models. Difference-in-differences is an interactive fixed effects model with panel data (Kropko and Kubinec 2020). Regression discontinuity design is a form of non-parametric regression where the predictor is evaluated at a single point (Lee 2008; Calonico, Cattaneo, and Titiunik 2014). Causal identification in these models requires knowing a fair amount of detail about the true data-generating process (i.e., the causal graph), leading some to argue that these methods have become over-employed and poorly understood (Kahn-Lang and Lang 2018; Caughey and Sekhon 2011; Grimmer 2011). To proxy for random assignment, scholars have come to emphasize statistical tests for incomparability between units, such as pre-treatment trends, but these statistical fixes can have the unfortunate side effect of encouraging type II errors given the low power of these tests (Roth and Sant'Anna 2021; Angrist et al. 2019).

These models are certainly useful, but are best understood within the general framework of observational analyses that can provide causal knowledge given the right conditions. Fitting observational methods into the framework of ignorability can certainly help elucidate some problems of causal inference, but I would argue that doing so is not strictly necessary for inference. One oft-cited example of the ideal conditions for observational inference comes from the analysis of the correlation between smoking and lung cancer (Cornfield et al. 1959). Another, although it is not often expressed this way, is the application of polling aggregation models to predict electoral outcomes (Campbell 2016). While election forecasting models are not perfect and sometimes over or under-predict, the underlying causal process is usually undisputed: if a person is asked who they will vote for the day before the election, they will very likely cast a vote for that person in the ballot box. In this situation, there is plenty of data, we can measure most of what we want to know,

and the measures are fairly direct of the underlying outcome.

4 Mechanistic Causation

While the previous two approaches are often contrasted as the only ways of thinking about causality in the sciences, there is a third philosophy that is gaining traction among qualitative researchers, especially in sociology and political science. There has been much work in these disciplines in recent years on establishing how case study research can identify mechanisms that link causal variables (Bennett 2014; Abbott 1992; Evera 1997). Again, instead of considering this line of inquiry to be a subsidiary issue to the question of causality, I propose that mechanisms are a part of what we mean by causality, and hence are their own silver standard. When ideal conditions are met, we can infer causality with reasonable, though never perfect, confidence.

Several definitions of mechanisms have been proposed in the literature (Mahoney 2012; Gerring 2017). I use for this paper the standard of Waldner (2015) that mechanisms are invariant processes connecting causal variables to each other. The example he uses is very instructive: the person who discovered the mechanisms through which aspirin relieved heart pressure, Sir John Vane, received a Nobel Prize in 1982, long after it had been conclusively established that aspirin had these lasting effects. This Nobel prize is puzzling under either the observational or experimental approaches to causality: if the strength of association was indisputable and random assignment had provided an estimate of all possible counterfactuals, then how could this person receive a Nobel prize for an entirely distinct discovery?

The answer is that when humans conceive of causality, we imagine there to be some kind of process linking cause and effect. Defining exactly what this is process is can be difficult, but the invariant standard is a helpful step. We are looking for processes that are so low-level that they operate similarly in all contexts. In a social scientific setting, we might think of basic emotions like fear, anger and happiness (Pearlman 2013), or the rational actor model (Elster 1994).

Waldner's theory of causation is particularly useful in this paper as it directly compares the mechanistic thinking of causation to Pearl's use of network diagrams for causal structures. To integrate mechanisms, we can introduce labels on the edges that identify which mechanism is in operation.[2] Importantly, *these mechanisms are not variables*, as causal mediation analysis pre-supposes (Imai, Keele, and Tingley 2010). Rather, mechanisms are root processes that are at the limit of our observational capacity and are so fundamental as to be nearly determinate to the outcome. Causal mediation assumes we can collect ample data on the mediators, but if we are at the limit of our observational capacity, the scope of data collection is necessarily limited and the necessary quantification may result in important loss of information that is difficult to capture in a rectangular dataset (Collier, Brady, and Seawright 2010). It is of course always possible to improve the quantified measurement of mechanisms to the point that they can be treated as a mediators, which is a point of connection between mechanistic and other modes of inference.

Like the other two silver standards, under ideal conditions mechanistic research can provide strong evidence of causality. In the social sciences this entails collecting exhaustive evidence on a person's decision-making in what has come to be called process-tracing. Process-tracing methodologists refer to the idea of a "smoking gun" as the kind of evidence that establishes causality within this framework, such as obtaining a private diary of an important leader that describes in detail why they made their decisions (Evera 1997; Collier, Brady, and Seawright 2010; Bennett 2014; Humphreys and Jacobs 2015). These ideal conditions are relatively rare, but like experimental and observational approaches, we can have confidence in inferring a high level of causal knowledge when we have all the information we might want about how X changed into Y.

Of course, qualitative inference does lend itself to single-act causation as opposed to establishing general trends between variables. However, to the extent that the variables under study are the same across conditions, then we could presume that our knowledge of the mechanisms linking these variables is similarly invariant. If we can establish what the mechanisms are then we can be

[2]While Waldner (2015) first proposed that edges could be labeled as representing mechanisms, Pearl has also come to the same interpretation. See Cinelli and Pearl (2019), footnote 2.

more confident that the variables are causally associated in addition to any relationship we estimate using observational or experimental statistics.

5 Synthesis

> "Of course it is happening inside your head, Harry, but why on earth should that mean that it is not real?"
>
> – Professor Albus Dumbledore from J.K. Rowling's *Harry Potter and the Deathly Hallows*

The point of presenting each of these ways of inferring causality is to promote a realist conceptualization of causal inference for the social sciences. To paraphrase the U.S. Supreme Court Justice Potter Stewart, we may not be able to define causality exhaustively, but we know it when we see it. In an exhaustive summary, Sloman and Lagnado (2015) argue based on extensive experimental evidence that while the network perspective of Pearl solved many issues in formalizing causal thinking, it still "has not offered a silver bullet that answers all questions about human thought" (p. 240). What causality exactly means is a subject for debate, even if we have made remarkable strides in terms of precisely detailing certain causal logics in the past three decades.

This latent source of uncertainty over causality is rarely discussed in the papers cited above regarding the study of causal processes. Rather, frameworks are presented as a way of defining causality, but over time the framework becomes synonymous with the term. In common parlance in the social sciences, a "causal inference" model implies either a randomized control trial or the usage of certain statistical methods that can be expressed using Rubin's potential outcomes notation. This slippage in terminology puts the cart before the horse: causality does not inherit from counterfactuals; rather, counterfactuals arose as a way of expressing what is meant by causality.

This realist point is not to suggest, as some have, that "causal pluralism" necessitates a refusal to admit standards for causal inference (Reiss 2009). Rather, the aim is for a unitary conception

of causation that can still allow for diversity in research methods (Gerring 2006). Over time, we have learned more about what we mean by the word causal, and continued research should help us more astutely understand the inferences we make and the reasons why we make them (Spirling and Stewart 2022).

6 Entropy as a Neutral Standard

It is difficult to properly conceptualize different modes of inference as these are relatively abstract categories. To make matters more concise, I next present the concept of entropy to formalize the idea that diverse research methods can provide varying amounts of causal knowledge, as opposed to delineating some methods as causal and others as observational. In general, entropy describes the decay of a system, such as gas molecules moving farther and farther apart to fill a sphere. Statistical, or Shannon, entropy applies the same concept to probability, providing a measure of the "information" in a random variable (Shannon 1948). In general, as probability distribution becomes more equal or uniform, entropy increases because all outcomes are equally likely, whereas when a probability distribution becomes more degenerate or peaked, entropy decreases as some outcomes are more certain than others.

Shannon entropy H is defined as a simple formula applied to a distribution of N probabilities that cumulatively sum to 1:

$$H = -\sum_{n=1}^{N} p_n \log p_n \tag{1}$$

The formula in (1) is unfortunately not intuitive, but its appeal is in meeting certain qualifications for determining an entropy measure of probability, including that it increases as it moves away from neutral probabilities over outcomes (such as $\frac{1}{N}$) and reaches a minimum at 0 when any of the probabilities are equal to 1. The units of entropy are determined by the type of logarithm employed. Because I am interested in entropy as a framework rather than with a particular empirical

application, I use an unconventional logarithmic base of 1.01:

$$H = -\sum_{n=1}^{N} p_n \log_{1.01} p_n \tag{2}$$

A base of 1.01 means that every unit increase in entropy equals a one percent increase in entropy. Figure 3 plots entropy calculations for probability distributions with varying levels of total uncertainty or spreadout-ness. What is important to note is that all of these distributions have the same expected, or average, value, but are nonetheless very different statements about underlying uncertainty. Roughly speaking, the uniform distribution has 100 percent more entropy than the normal distribution, which has 30 percent more entropy than the student's T and Laplace distributions. These plots show why entropy is a powerful heuristic: it captures our sense of how certain we are of the empirical possibilities underlying a distribution of probability that is independent of the form of the distribution. If we know nothing about a process, we can assume a uniform distribution which leaves probability mass on all possible outcomes. But if we know more about how a process operates, we can considerably reduce our uncertainty (and hence entropy) by choosing a more specified distribution.

While entropy has been applied successfully to many statistical problems, my intention in defining it here is to think of it as a way to understand the relative value of the causal paradigms previously discussed. Ultimately, the goal of the social sciences should be to reduce entropy whenever possible in terms of our understanding of how the social world operates. If we have more certain knowledge of the distribution of outcomes, we can state with reasonable confidence that our knowledge is increasing (Gerring 2006). To do so, we have to produce new propositions that explain human behavior and allow us to make judgments about what is more or less likely to occur.

The maximum entropy principle provides further clarity about how we can maximize knowledge while avoiding over-confidence. Jaynes (2003) defines the maximum entropy principle as always preferring a distribution of higher entropy conditional on including all known facts in the distribution. For example, suppose we wanted to predict stock market prices. Lacking any special

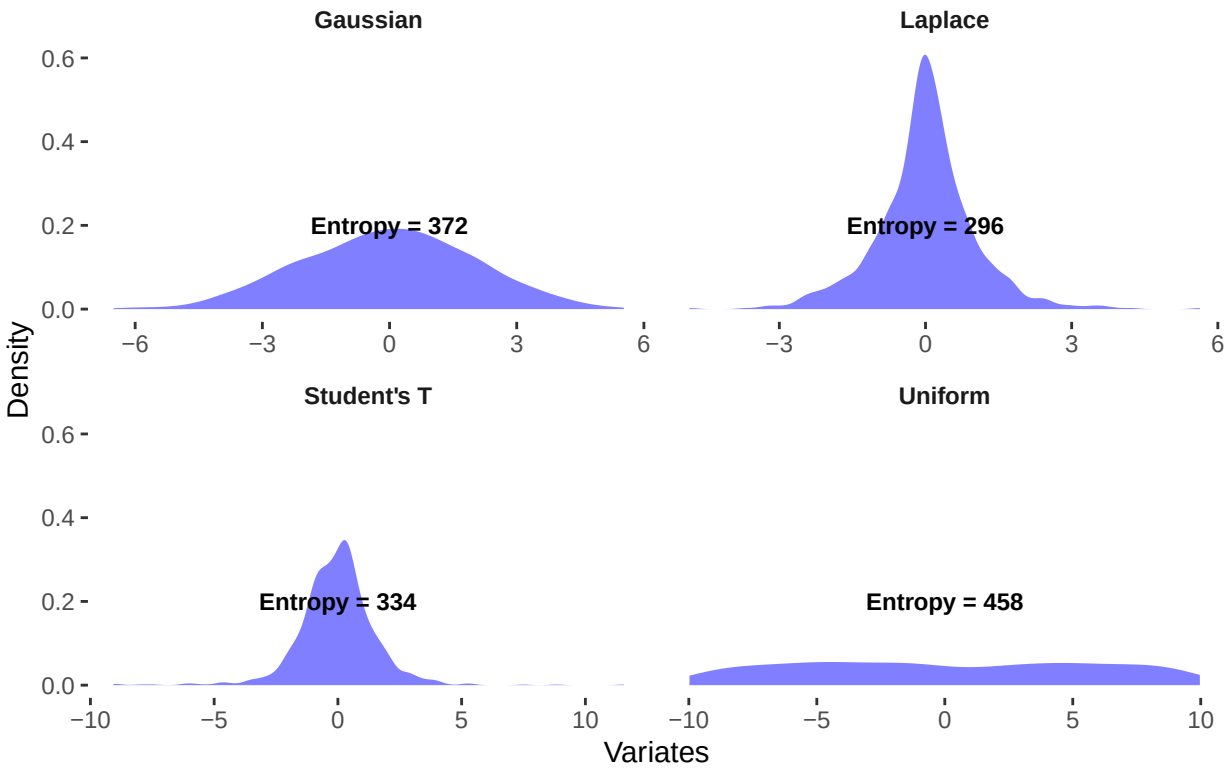

Figure 3: Entropy Calculations Based on Empirical Densities of Statistical Distributions

knowledge into stock prices, we would want our uncertainty to reflect the fact that all we have to analyze are the movements of individual stocks over time–we would want to maximize entropy, or uncertainty, given the data we have. But if we knew that the Federal Reserve intended to increase interest rates, we could include that information in our model and consequently obtain a lower entropy distribution.

In other words, we want to learn new facts about the world such that we reduce our entropy in understanding causal relations. At the same time, we want to maximize entropy given what we know to reduce blind spots and over-confidence. Causal inference involves striking this delicate balance between assuming too much and assuming too little.

This framework helps resolve some inconsistencies in how models are incorporated in the social sciences. On the one hand, more complex models are seen as embodying increased knowledge (Clarke and Primo 2007; Slough 2019). On the other hand, there has been considerable push back at models that appear to be baroque and less easy to explain than tried-and-true ordinary least squares (OLS) regression (Angrist and Pischke 2008). Maximum entropy helps explain these mixed feelings: we should prefer more complex models over simple models because our over-arching aim should be to reduce entropy, and more complex models have less entropy. On the other hand, we do not want to reduce entropy without a good reason lest we over-state our certainty (Frank 2009).

To use entropy to understand causal paradigms, I return to Pearl's causal diagrams. My intention is not to suggest that Pearl's theory is the final take on causality, but rather that network relations are deeply intuitive representations of human thought patterns. As such, they are a helpful starting block for comparing very different representations of causality. There are existing applications of entropy to causal graphs, but the aim in the literature is to uncover hidden confounders given a set of observed data as opposed to making larger statements about research design (Kocaoglu et al. 2020; Tee, Parisis, and Wakeman 2016; Wieczorek and Roth 2019).

Let us imagine that we launched ourselves in a spaceship and landed in a foreign world. We

have almost no prior knowledge of how people on this planet relate to each other, but we want to understand how the world's residents select their leaders. All we can do is come up with a list of plausible factors that might affect leader selection. Given our experience of such occurrences on earth, we come up with the following list of variables: political ideology (I), economic benefits (E), ethnic affinity (A), leader personal qualities (Q), and the risk of conflict (C). We can think of these variables as nodes in a network all connected with the outcome of leader selection (Y) as is shown in Figure 4.

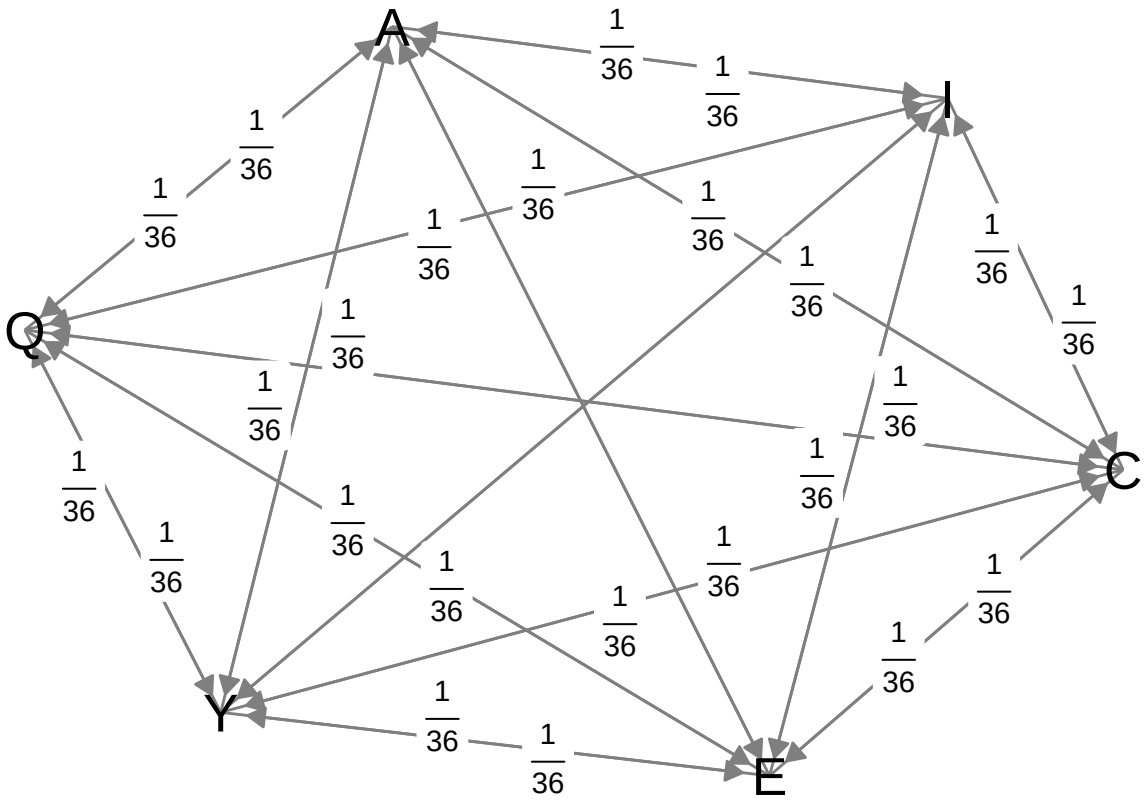

Figure 4: Causal Diagram with Complete Uncertainty

Each edge in this graph is labeled with the dual expected probabilities that a link exists in either direction, with each one labeled $\frac{1}{36}$ to represent our current ignorance, i.e., any link between any of the nodes is equally likely in either causal direction. The edges in this network are bi-directional to show that our in our state of complete uncertainty we cannot even say what the direction of causality could be. While the uncertainty in this figure is extreme, it comes closer to the actual state

of social science research than the elegant causal graphs in many texts derived from mechanical examples.

Given the previous discussion, the question now is to reason about which method of causal inference to apply to Figure 4. The easiest way to answer this question, and one often chosen, is simply to choose whichever method best fits the researcher's skills and experiences. While very practical, it poses a chicken-and-egg problem, and only shifts the question to which paradigm researchers should invest in to gain experience and skills.

I propose that a better heuristic is to ask what would reduce entropy in the causal graph. There are very complicated entropy statistics for graphs, but to simplify matters I apply the entropy formula to each edge probability in Figure 4:

$$-\sum_{1}^{36} \frac{1}{36} \log_{1.01} \frac{1}{36} = 360$$

We start with the considerably high number of 360. At this amount of entropy, we are not likely to be wrong, but we also cannot say much of value about our study of this foreign planet's society. Initially, let us consider a choice between an experimental and an observational analysis. Suppose than with an experiment we can determine directly the probability of the connection between ethnic affinity (A) and leader selection (Y).[3] If we pull off a quality experiment, we can double the probability of the link from $A \rightarrow Y$ and reduce the probability of the link from $Y \rightarrow A$ (reverse causality) to 0. We can consider the experiment to be *causally identified* because it is possible, given enough experimental data, to either rule out or establish the relationship between these two nodes (Keele 2015). Then we can re-calculate our entropy measure, which shows a decrease in entropy of 4 percent:

$$-[(\sum_{1}^{34} \frac{1}{36} \log_{1.01} \frac{1}{36}) + \frac{2}{36} \log_{1.01} \frac{2}{36}] = 356$$

[3]We ignore for the time being the difficulty in enacting these research designs.

However, suppose that if we conducted an observational data analysis, we could increase or decrease the probability of the links between leader selection and economic benefits (E), leadership quality (Q), ideology (I), ethnic affinity (A) and conflict (C) from $\frac{1}{36}$ to $\frac{1.5}{36}$ while lowering the opposite links to $\frac{0.5}{36}$. This research design is not causally identified because we cannot be sure, i.e. we cannot know with probabilities approaching one, what effect any one of these variables have on the outcome, *only their joint distribution*. This analysis would result in the following change in entropy:

$$-[\sum_{1}^{26} \frac{1}{36} \log_{1.01} \frac{1}{36} + \sum_{1}^{5} \frac{1.5}{36} \log_{1.01} \frac{1.5}{36} + \sum_{1}^{5} \frac{0.5}{36} \log_{1.01} \frac{0.5}{36}] = 356$$

In other words, in this example, the observational and experimental studies would have similar effects on reducing the entropy of the total system *even though they made very different statements about the underlying causal structure*. Intuitively, we can learn a lot from establishing a specific causal link in a specific direction with a high degree of certainty, but we can also learn a lot from examining associations between variables, even if we cannot arrive at conclusive predictions. The point of this exercise is not to suggest that observational methods are better than experimental methods, but rather that the value of each depends on the nature of the causal problem, and it is *not* always the case that experiments produce more causal knowledge than observational studies. From a Bayesian point of view, we could re-state this problem as meaning that we should always prefer the research design that increases our knowledge relative to our prior, even if the knowledge we obtain has residual bias (Little and Pepinsky 2018).

I now formally define the causal entropy measure as the Shannon entropy $H(\cdot)$ of an N ordered set of variables $\{x_1, ...x_N\}$ that can be represented by a causal graph V which meets all of Pearl's requirements: it includes a set of directed edges $e \in E$ between variables in the graph, and is acyclic. We can then take the Shannon entropy $H(\cdot)$ of the joint distribution of these variables, which can be denoted $P(v)$, following Pearl's notation:

$$H(P(v)) = -\sum_{i=1}^{N} P(x_i|pa_i)P(pa_i)\log P(x_i|pa_i)P(pa_i) \tag{3}$$

Where the notation $P(x_i|pa_i)$ indicates that each component of $P(v)$ is the conditional distribution of each variable x_i in V with respect to the set of its ancestors pa_i that are causally relevant to x_i. This formalism captures the procedure done earlier in which probabilities are assigned to each directed relation in a causal graph. Because $P(v)$ is a joint distribution over all such relations, it meets the requirement that the probabilities of all of the causal relations sum to 1.

Formally, we can then consider research designs as representing different joint distributions, such as $P(v)$ and $P'(v)$. We can state that a research design that results in $P(v)$ creates less causal knowledge than a research design that produces $P'(v)$ iff:

$$H(P(v)) > H(P'(v)) \tag{4}$$

The one complication is that entropy is only defined over discrete variables. However, it is straightforward to calculate entropy of a continuous variable through a binning procedure, and the measure is available via a wide array of statistical software packages (Hausser and Strimmer 2021).

As mentioned previously, we can consider a distribution of mechanisms in the causal graph V, which I denote as the set Ω. Each directed relation $P(x_i|pa_i)$ would have a corresponding distribution over mechanisms Ω, where $P(\Omega|x_i, pa_i)$ represents the joint distribution of all mechanisms for a given relation. Because of causal graphs' Markovian property (Pearl 2000, Ch. 1), we can simplify the expression to $P(\Omega|x_i)P(pa_i)$ because each variable in a causal graph is independent of other variables in the causal graph once we factor in its parents. In other words, we only need to consider the mechanisms of each node but not each preceding node. We can then consider the joint of both distributions:

$$H(P(v, \Omega)) = -\sum_{i=1}^{N} P(x_i|pa_i)P(\Omega|x_i)P(pa_i)\log P(x_i|pa_i)P(\Omega|x_i)P(pa_i) \tag{5}$$

If we consider a graph V that had identical distributions for the probabilities of causal relations $P(v)$, but different distributions for mechanisms $P(\Omega')$, we could then define when a study that increased our understanding of mechanisms would be preferable:

$$H(P(v, \Omega)) > H(P(v, \Omega')) \tag{6}$$

So far I have shown all results using Shannon entropy H. It is also possible to examine the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) between different probability distributions of $P(v)$. This measure of *relative* entropy is quite similar to Shannon entropy, so I do not address this further. Using the KL divergence could be useful in a situation where comparisons are made between multiple competing research designs rather than just two or three, and the aim is to compare distances across probability distributions.

The intention of this formulation is to show that the concept of statistical entropy sheds light on the difficult decisions that must be made when considering research designs. Although there may be times when directly calculating changes in entropy from causal graphs is necessary, the formalism can also be used as a heuristic for evaluating research designs without having joint probability distributions. Ultimately, the criterion of reducing entropy suggests that we aim for maximizing the amount we can learn, i.e. causal knowledge, from an application of any of the paradigms. Even weakly causally-identified designs can contribute causal knowledge. To illustrate the principle further, I apply the theory to an important recent area of research.

6.1 Case Study: COVID-19 Vaccines

The outbreak of the COVID-19 pandemic offers an important test case for understanding how researchers employed research designs to understand and prevent COVID-19 infections. Because of the speed of the outbreak and the enormous scale of research efforts, there was relatively little time for traditional disciplinary norms to determine research designs. The SARS-CoV-2 virus did

not care for disciplinary preferences, forcing researchers to employ whatever methods they had available to study the pandemic. The extreme pressures of this exogenous shock is the reason why I choose this particular area of research even though it is not a part of the social sciences proper. Under pressure, scientists went with their causal intuitions as opposed to relying on traditional disciplinary hierarchies, producing innovative work that built across modes of inference rather than relying on one mode at the expense of others. In addition, the massive levels of funding available from governments overcame one of the most common non-methodological factors in determining research designs, permitting issues of inference to become relatively more important.

Although there are many possible research questions, in this case study I focus on one crucial area: the development of vaccines. At first blush, it would seem that vaccines are a relatively straightforward exercise in terms of research design. After months of development, the drug companies Pfizer-BioNTech and Moderna released studies describing massive RCTs employing hundreds of thousands of volunteers over months. These studies provided concise and clear numbers concerning *vaccine efficacy*, or the ratio of infected individuals in the treatment group to the number infected in the control group (Polack et al. 2020; Baden et al. 2021). Because the control group never received a vaccine, the difference between the two groups could be directly attributed to the drug.

In causal terms, this RCT solved a difficult yet well-known problem in studying COVID-19 infections: those who voluntarily participate in a COVID-19 vaccination study could be either more or less likely to be infected compared to those who would not want to volunteer for a vaccine. To give just one example, younger individuals showed less interest in vaccines compared to older individuals, and were also much less likely to become severely ill. On the other hand, younger people may have been more likely than older people to contract a COVID-19 infection because they had less fear of serious illness or death. As a result, any naive comparison of a group of volunteers and the general population could end up conflating age differences with vaccine uptake (Hodgson et al. 2021; Baack et al. 2021). This causal identification problem is a straightforward example of confounding, as shown in Figure 1. Any variable which could explain both vaccine uptake and

the incidence of COVID-19 would be a confounding variable, and without confidence that we can collect data on and measure all confounding variables, we may not be able to identify the direct relationship between the vaccination and efficacy.

As is well-known, the Pfizer-BioNTech and Moderna trials proved to be a paragon of RCT methods, showing remarkably strong effects of the vaccine on efficacy, above 90%. At this point in the narrative, it would seem that RCTs had proven themselves as the gold standard: we had established that the vaccines worked, and now we could move forward with ending the pandemic. Indeed, such sentiments were common when the vaccines were introduced, leading to a relaxation of restrictions in the summer of 2021 (Bauer et al. 2021; Tregoning et al. 2021).

Fairly quickly, however, it became evident that the RCTs themselves were not sufficient to answer all the questions about vaccine efficacy. There were two main problems: first, people wanted to know how the vaccine performed *in the population*, which required an attention to the confounding variables that the RCT successfully ignored (Hungerford and Cunliffe 2021). Second, the arrival of vaccine variants forced a re-evaluation of the vaccines' efficacy as RCT trials could not be run fast enough to keep up with new variants (Andrews et al. 2022). These issues required both observational and mechanism-based modes of inference, as I will explicate below.

Table 1: $Pr(I|T, Z = \text{Old})$

	I = Infected	I = Not Infected
T = Vaccine	.15	.85
T = No Vaccine	.85	.15

We can examine the entropy reductions of these different interventions on the causal graph by detailing the causal variables' conditional probability distributions. For simplicity, I will take as my starting point the causal graph in Figure 1, where the nodes can be relabeled so that there is one treatment variable, V for vaccine, one outcome, I for infection, and a confounder Z , which I will consider to be age. To simplify matters, I will treat each variable as having two discrete

values. Tables 1, 2 and 3 show plausible marginal conditional probability distributions for the three variables. To analyze the relationships in the causal graph, we need to consider two conditional probabilities, $Pr(I|V, Z)$ and $Pr(T|Z)$. Because the first conditional probability involves two conditioning variables V and Z, I separate this distribution of I into two separate tables for Young and Old subjects as can be seen in Tables 1 and 2.

Table 2: $Pr(I|T, Z = \text{Young})$

	I = Infected	I = Not Infected
T = Vaccine	.02	.98
T = No Vaccine	.98	.02

Table 3: $Pr(T|Z)$

	T = Vaccine	T = No Vaccine
Z = Young	.1	.9
Z = Old	.9	.1

I assume here that these are the true probabilities of treatment efficacy and the confounding effect of age on vaccine uptake. It is important to note, too, that age also affects vaccine efficacy, with the vaccine more efficacious among the young than the old, as studies have shown (Bell and Kutzler 2022). To calculate entropy, we will need to start with a prior distribution representing what we think these relationships could be, or what I will call the null graph. For simplicity, I will assume a uniform prior for the null graph, i.e., that all of the probabilities in the tables are equal to exactly 0.5. While not shown, I can calculate the entropy by considering the full joint distribution of the causal graph, which involves creating a much larger table for all values of Z, T and I, i.e. $P(I, Z, T) = P(I|T, Z)P(T|Z)P(Z)$. For reference, I also assume that the population is 25% young and 75% old.

Calculation of Shannon entropy (not shown) indicates that the null prior graph has an entropy of 209, while the true graph has an entropy of 124. These two figures give us the relative space within which we can plausibly learn about this outcome. If we end up with an entropy of less than 124, we will be over-confident, inferring causality to what are in fact random events. Respecting the lower bound reflects the principle of maximum entropy discussed earlier: we should not want to be more certain of conclusions than the underlying causal process permits.

We can directly calculate the reduction in entropy for the experimental analysis of the vaccine by inserting values of the conditional probability distributions from the true graph for $P(I|T, Z)$ into the null graph. By adding in the true values for the conditional distribution, not just the average treatment effect, I assume that the treatment was high-powered enough to inform us about the true joint distribution of both the treatment and the potential confounder, age, as seemed to be true for most of the vaccine trials with tens of thousands of subjects enrolled. The entropy of this high-powered experimental analysis is 143, which is only 19% larger than the true entropy. As such, the experimental technique was clearly a powerful way of learning about this causal process.

However, the analysis left important information unanswered on the causal graph, in particular what the relationship is between vaccine uptake and age. While the relative entropy would seem small, it is still an appreciable amount, and when vaccines were deployed to the population, it became a crucial factor for understanding the success of the vaccines (Hodgson et al. 2021). If sicker people were more likely to take the vaccine, then the measures of vaccine efficacy from the total population would understate the efficacy of the vaccine. Understanding how the vaccine interacted with population demographics required obtaining data about vaccine uptake in the "real world" (Chodick et al. 2021), especially to combat misinformation about the efficacy of the vaccine by anti-vaccination groups. For this reason, it is clear that even though the most important question about the vaccine was answered by an RCT, there was ample room for observational studies that collected data on the spread of the vaccine and relative rates of COVID-19 incidence in the population.

Ultimately, these observational studies were necessary to uncover the remaining entropy in the causal graph, equivalent to a 19% reduction in entropy. While this reduction was not in large as the reduction due to the experiment, it is important to note that this reduction could not be obtained from the experiment itself as it involved fixing the vaccine node V to a particular value, such as with Pearl's do operator. As argued previously, causal identification is sufficient for causal knowledge to be obtained, but it is not necessary. In this case, causal identification by fixing V to a specific value in an RCT prevented any analysis of the $P(T|Z)$ relationship because by definition it removed that causal arrow from the graph. For this reason, an observational analysis that established the $P(T|Z)$ relationship—varying vaccine interest by age—would likely be labeled as descriptive, not causal. However, this distinction is relatively arbitrary when we consider the causal graph holistically using a measure like entropy.

Finally, it is important to note that mechanistic analysis also played an important role in determining the efficacy of vaccines in the pandemic. As mentioned earlier, the success of the vaccines waned depending on the mutations of the virus, and it was infeasible to keep running large RCTs for each variant. This is a kind of threat to inference that is rarely discussed, and could be described as "temporal validity" (Munger 2019). To address this problem, scholars examined whether the same mechanism underlying the vaccine's efficacy also occurred in the same way with variants, namely, the production of virus-neutralizing antibodies. These studies were not necessarily statistical in nature, involving close examination of relatively small numbers of petri dishes with new SARS-CoV-2 variants in blood that had vaccine-induced antibodies (Yadav et al. 2021; Hoffmann et al. 2022). Furthermore, these studies could not be directly integrated into the causal graph examined above because they refer to factors that are not present on the graph itself, i.e., minuscule changes in antibody levels.

We can expand the analysis to incorporate the antibody mechanism if we give it two values, High and Low. With these two values in the set Ω, we can then calculate the entropy of the combined causal graph conditional on this mechanism for the relationship between vaccine V and infections I:

$$H(P(I, Z, T, \Omega)) = H(P(I|T, Z)P(\Omega|T)P(T|Z)Pr(Z)) \tag{7}$$

If we start with a null graph where the probability of Ω = High is 0.50, and the true value is 0.9, then we have null and true entropy values of 279 and 156 respectively. If we performed the experiment successfully but without learning about mechanisms, we would obtain an entropy value of 198 while an experiment that also involved learning about mechanisms would result in an entropy of 161, or 37% less. As can be seen, even with a simple binary mediator, relatively large reductions in entropy are possible by obtaining evidence of its true value. The reason for this large reduction is due to the fact that we know that the mechanism must be present for the causal story to hold about immunological response. If we were less certain about whether these mechanisms mattered in this case, we would not see as large reductions.

Of course, it could always be possible to look at antibodies as a mediator and expand the causal graph, allowing us to use mediation analysis to formally test for whether antibodies mediate the vaccine (Imai, Keele, and Tingley 2010; VanderWeele 2016). What is important, however, is that statistical tests were not necessary to make evidentiary statements about the presence or absence of the proposed mechanism. The proximity of observation and the logical necessity of antibody levels changing helped obtain causal inference about the performance of the vaccine against new variants even if there was no uncertainty interval or p-value attached. Statistical methods may not work well in this mode of inference because the attention to micro-processes often entails serious limitations in data collection in favor of richly textured information (Collier, Brady, and Seawright 2010). The use of priors can permit a form of Bayesian inference, though it is technically more of a logical approach based on probability theory than a statistical method per se (Humphreys and Jacobs 2015). In any case, assigning a probability value to the mechanism Ω ultimately involves the researchers' qualitative judgment, not data collection. By maximizing knowledge of mechanisms, researchers were able to warn about dangerous variants long before either observational or experimental studies could be completed.

The point of this case study was not to argue that one mode of inference was superior to the other, but rather how each kind–experimental, observational and mechanistic–played a valuable role in causal learning about the efficacy of the vaccine. At different stages in the pandemic, each of these modes of inference helped determine the relative vaccine efficacy both against variants and in the population as a whole. While the RCT achieved the highest reduction in entropy, other modes of inference had an important role to play. Once cost-benefit factors are included, such as the need to learn about efficacy against new variants for the purposes of booster shots, smaller overall changes in entropy can still be very important from a social welfare perspective.

7 Conclusion

Ongoing debates about causal inference threaten to create divides that impede research progress. Part of the problem is the growing assumption that causal inference requires an RCT or at minimum a model expressed in terms of potential outcomes. This divide separates research into causal and "mere" association, with the former preferred over the latter without reference to the relative amount of causal knowledge to be gained.

Rather than point to problems with RCTs as a reason to distrust them, I aver that the underlying issue is that we conceive of causality as a binary process: either a piece of research is or is not causal. It is not that RCTs have more issues than are commonly acknowledged, but rather that the definition of RCT as the gold standard means that we artificially deflate the value of other modes of causal analysis. We should take a charitable approach by admitting that various theories and practices of inference contain varying amounts of causal knowledge, with the amount varying in relation to the credibility of the research design. However, as the maximum entropy principle shows, we cannot evaluate the relative causal knowledge obtained without a consideration of what is and is not known about a given causal system.

The principle of entropy provides one helpful framework by imagining the benefit of a study from

the relative reduction in entropy it achieves. By using such a continuous criterion, we can admit that we can learn from studies that are not causally identified without lowering standards of inference. This framework can help guide both qualitative and quantitative decisions about formulating research inquiries, depending on the level of formalization needed.

References

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72 (1): 1–19.

Abbott, Andrew. 1992. "From Causes to Events: Notes on Narrative Positivism." *Sociological Methods and Research* 20: 428–55.

Altman, Naomi, and Martin Krzywinski. 2015. "Association, Correlation and Causation." *Nature Methods* 12 (10): 899–900. https://doi.org/10.1038/nmeth.3587.

Amrhein, Valentin, Sander Greenland, and Blake McShane. 2019. "Scientists Rise up Against Statistical Significance." *Nature*, March.

Andrews, Nick, Julia Stowe, Freja Kirsebom, Samuel Toffa, Tim Rickeard, Eileen Gallagher, Charlotte Gower, et al. 2022. "Covid-19 Vaccine Effectiveness Against the Omicron (b.1.1.529) Variant." *New England Journal of Medicine* 386 (16): 1532–46. https://doi.org/10.1056/NEJMoa2119451.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides' Rule Redux." *American Economic Review* 1 (3): 309–24.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

———. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Baack, Brittney N., Neetu Abad, David Yankey, Katherine E. Kahn, Hilda Razzaghi, Kathryn Brookmeyer, Jessica Kolis, Elisabeth Wilhelm, Kimberly H. Nguyen, and James A. Singleton.

2021. "COVID-19 Vaccination Coverage and Intent Among Adults Aged 18–39 Years — United States, March–may 2021." *Morbidity and Mortality Weekly Report* 70 (25): 928–33. https://doi.org/10.15585/mmwr.mm7025e2.

Baden, Lindsey R., Hana M. El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, et al. 2021. "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine." *New England Journal of Medicine* 384 (5): 403–16. https://doi.org/10.1056/NEJMoa2035389.

Bauer, Simon, Sebastian Contreras, Jonas Dehning, Matthias Linden, Emil Iftekhar, Sebastian B. Mohr, Alvaro Olivera-Nappa, and Viola Priesemann. 2021. "Relaxing Restrictions at the Pace of Vaccination Increases Freedom and Guards Against Further COVID-19 Waves." *PLOS Computational Biology* 17 (9): e1009288. https://doi.org/10.1371/journal.pcbi.1009288.

Beck, Nathaniel. 2006. "Is Causal-Process Tracing an Oxymoron?" *Political Analysis* 14 (3): 347–52.

Bell, Matthew R., and Michele A. Kutzler. 2022. "An Old Problem with New Solutions: Strategies to Improve Vaccine Efficacy in the Elderly." *Advanced Drug Delivery Reviews* 183 (April): 114175. https://doi.org/10.1016/j.addr.2022.114175.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2017. "Redefine Statistical Significance." *Nature Human Behavior* 2: 6–10.

Bennett, Andrew. 2014. "Disciplining Our Conjectures: Systematizing Process Tracing with Bayesian Analysis." In, edited by Andrew Bennet and Jeffrey T. Checkel, 276–98. Cambridge Univ Press: Cambridge, UK.

Bethe, H. A. 1990. "Supernova Mechanisms." *Reviews of Modern Physics* 62 (4): 801–66.

Boyko, Edward J., and Beth W. Alderman. 1990. "The Use of Risk Factors in Medical Diagnosis: Opportunities and Cautions." *Journal of Clinical Epidemiology* 43 (9): 851–58.

Brashier, Nadia M., and Elizabeth J. Marsh. 2020. "Judging Truth." *Annual Review of Psychology*, no. 71: 499–515.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Con-

fidence Intervals for Regression Discontinuity Designs." *Econometrica* 82 (6): 2295–36.

Campbell, James E. 2016. "Forecasting the 2016 American National Elections." *PS: Political Science & Politics* 49 (4): 649–54.

Caughey, Devin, and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close u.s. House Races, 1942-2008." *Political Analysis* 19 (4): 385–408.

Chodick, Gabriel, Lilac Tene, Tal Patalon, Sivan Gazit, Amir Ben Tov, Dani Cohen, and Khitam Muhsen. 2021. "Assessment of Effectiveness of 1 Dose of BNT162b2 Vaccine for SARS-CoV-2 Infection 13 to 24 Days After Immunization." *JAMA Network Open* 4 (6): e2115985. https://doi.org/10.1001/jamanetworkopen.2021.15985.

Cinelli, Carlos, and Judea Pearl. 2019. "Generalizing Experimental Results by Leveraging Knowledge of Mechanisms." *Technical Report*. http://ftp.cs.ucla.edu/pub/stat_ser/r492.pdf.

Clarke, Kevin A., and David M. Primo. 2007. "Modernizing Political Science: A Model-Based Approach." *Perspectives on Politics* 5 (4): 741–53.

Collier, David, Henry E. Brady, and Jason Seawright. 2010. "Toward an Alternative View of Methodology: Sources of Leverage in Causal Inference." In, edited by Henry E. Brady and David Collier. Lanham, MD: Rowman & Littlefield.

Cornfield, Jerfome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of the National Cancer Institute* 22 (1): 173–203.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Control Trials." *Social Science & Medicine* 210: 2–21.

Elster, Jon. 1994. "The Nature and Scope of Rational-Choice Explanation." In, edited by Michael Martin and Lee C. McIntyre, 311–22. Massachusetts Institute of Technology: Boston, MA.

Evera, Stephen Van. 1997. *Guide to Method for Students of Political Science*. Ithaca: Cornell University Press.

Fisher, R. A. 1935. *The Design of Experiments*. Oxford, England: Oliver; Boyd.

Frank, Steven A. 2009. "The Common Patterns of Nature." *Journal of Evolutionary Biology* 22

(8): 1563–85.

Gelman, Andrew. 2018. “Benefits and Limitations of Randomized Control Trials: A Commentary on Deaton and Cartwright.” *Social Science & Medicine* 210: 48–49.

Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No “Fishing Expedition" or “p-Hacking" and the Research Hypothesis Was Posited Ahead of Time.” http://www.stat.columbia.edu/ gelman/research/unpublished/phacking.pdf.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2014. “The Illusion of Learning from Observational Research.” In, edited by Dawn Langan Teele. Yale University.

Gerring, John. 2006. “Causation: A Unified Framework for the Social Sciences.” *Journal of Theoretical Politics* 17 (2): 163–98.

———. 2017. “Qualitative Methods.” *Annual Review of Political Science* 20: 15–36.

Gershman, Boris, David P. Guo, and Issa J. Dahabreh. 2018. “Using Observational Data for Personalized Medicine When Clinical Trial Evidence Is Limited.” *Fertility and Sterility* 109 (6): 946–51.

Gerstein, Hertzel G., John McMurray, and Rury R. Holman. 2019. “Real-World Studies No Substitute for RCTs in Establishing Efficacy.” *The Lancet* 393 (10168): 210–11.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2017. “Broken or Fixed Effects?” http://www.jcsuarez.com/Files/Suarez%3Csub%3ES%3C/sub%3Eerrato-BFE.pdf.

Goodman, Steven N., Daniele Fanneli, and John P. A. Ioannidis. 2016. “What Does Research Reproducibility Really Mean?” *Science Translational Medecine* 8 (341): 341–47.

Green, Donald P., and Alan S. Gerber. 2003. “Reclaiming the Experimental Tradition in Political Science.” In, edited by Ira Katznelson and Helen V. Milner, 805–33.

Grimmer, Justin. 2011. “An Introduction to Bayesian Inference via Variational Approximation.” *Political Analysis* 19 (1): 32–47.

Hausser, Jean, and Strimmer. 2021. *Entropy: Estimation of Entropy, Mutual Information and*

Related Quantities. https://CRAN.R-project.org/package=entropy.

Hernán, Miguel A. 2018. "The c-Word: Scientific Euphemisms Do Not Improve Causal Inference from Observational Data." *American Journal of Public Health* 108 (5): 616–19. https://doi.org/10.2105/AJPH.2018.304337.

Hodgson, Susanne H, Kushal Mansatta, Garry Mallett, Victoria Harris, Katherine R W Emary, and Andrew J Pollard. 2021. "What Defines an Efficacious COVID-19 Vaccine? A Review of the Challenges Assessing the Clinical Efficacy of Vaccines Against SARS-CoV-2." *The Lancet Infectious Diseases* 21 (2): e26–35. https://doi.org/10.1016/S1473-3099(20)30773-8.

Hoffmann, Markus, Nadine Krüger, Sebastian Schulz, Anne Cossmann, Cheila Rocha, Amy Kempf, Inga Nehlmeier, et al. 2022. "The Omicron Variant Is Highly Resistant Against Antibody-Mediated Neutralization: Implications for Control of the COVID-19 Pandemic." *Cell* 185 (3): 447–456.e11. https://doi.org/10.1016/j.cell.2021.12.032.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Humphreys, Macartan, and Alan Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109 (4): 653–73.

Hungerford, Daniel, and Nigel A. Cunliffe. 2021. "Real World Effectiveness of Covid-19 Vaccines." *BMJ* 374 (August): n2034. https://doi.org/10.1136/bmj.n2034.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4).

Imbens, Guido. 2018. "Understanding and Misunderstanding Randomized Control Trials: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 50–52.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Ioannidis, John P. A. 2018. "Randomized Control Trials: Often Flawed, Mostly Useless, Clearly Indispensable: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 53–56.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Kahn-Lang, Ariella, and Kevin Lang. 2018. "The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications." https://www.nber.org/papers/w24857.pdf.

Keele, Luke. 2015. "The Statistics of Causal Inference: A View from Political Methodology." *Political Analysis* 23 (3): 313–35.

Kocaoglu, Murat, Sanjay Shakkottai, Alexandros G. Dimakis, Constantine Caramanis, and Sriram Vishwanath. 2020. "Applications of Common Entropy for Causal Inference." In, 1751417525. NIPS'20. Red Hook, NY, USA: Curran Associates Inc.

Koslowski, Barbara. 1996. *Theory and Evidence: The Development of Scientific Reasoning*. Learning, Development, and Conceptual Change. Cambridge, MA, USA: A Bradford Book.

Kropko, Jonathan, and Robert Kubinec. 2020. "Interpretation and Identification of Within-Unit and Cross-Sectional Variation in Panel Data Models." *PLOS One* 15 (4): e0231349.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22 (1): 79–86. https://doi.org/10.1214/aoms/1177729694.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in u.s. House Elections." *Journal of Econometrics* 142: 675–97.

Levitt, Steven D., and John A. List. 2008. "Field Experiments in Economics: The Past, the Present, and the Future." https://www.nber.org/papers/w14356.

Little, Andrew T., and Thomas B. Pepinsky. 2018. "Learning from Biased Research Designs." *Social Science Research Network (SSRN)*, August. https://papers.ssrn.com/sol3/papers.cfm?abstract%3Csub%3Ei%3C/sub%3Ed=3236815.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41 (4): 570–97.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Munger, Kevin. 2019. "The Limited Value of Non-Replicable Field Experiments in Contexts

With Low Temporal Validity." *Social Media + Society* 5 (3): 2056305119859294. https://doi.org/10.1177/2056305119859294.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115 (11): 2600–2606.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science*.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

———. 2018. *The Book of Why: The New Science of Cause and Effect*. Penguin RandomHouse UK.

Pearlman, Wendy. 2013. "Emotions and the Microfoundations of the Arab Uprisings." *Perspectives on Politics* 11 (02): 387409.

Plümper, Thomas, and Vera Troeger. 2019. "Not so Harmless After All: The Fixed-Effects Model." *Political Analysis* 27 (1): 21–45.

Polack, Fernando P., Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, et al. 2020. "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine." *New England Journal of Medicine* 383 (27): 2603–15. https://doi.org/10.1056/NEJMoa2034577.

Przeworski, Adam. 2009. "Is the Science of Comparative Politics Possible?" In, edited by Carles Boix and Susan C. Stokes, 147–71. Oxford, UK: Oxford University Press.

Reiss, Julian. 2009. "Causation in the Social Sciences: Evidence, Inference, and Purpose." *Philosophy of the Social Sciences* 39 (1).

Roth, Jonathan, and Pedro H. C. Sant'Anna. 2021. "When Is Parallel Trends Sensitive to Functional Form?" *arXiv:2010.04814 [Econ, Stat]*, January. http://arxiv.org/abs/2010.04814.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66: 688–701.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *The Journal of Politics* 78

(3): 941–55.

Sampson, Robert J. 2018. "After the Experimental Turn: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 67–69.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.

Sloman, Steven A., and David Lagnado. 2015. "Causality in Thought." *Annual Review of Psychology* 66: 223–47.

Slough, Tara. 2019. "On Theory and Indentification: When and Why We Need Theory for Identification." *Working Paper*. http://taraslough.com/assets/pdf/theory_id.pdf.

Smith, Robert W. 1989. "The Cambridge Network in Action: The Discovery of Neptune." *Isis* 80 (3): 395–422.

Spirling, Arthur, and Brandon M. Stewart. 2022. "What Good Is a Regression? Inference to the Best Explanation and the Practice of Political Science Research." *Working Paper*.

Tee, Phil, George Parisis, and Ian Wakeman. 2016. "NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium." In, 1049–54. https://doi.org/10.1109/NOMS.2016.7502959.

Thapa, Deependra K., Denis C. Visentin, Glenn E. Hunt, Roger Watson, and Michelle Cleary. 2020. "Being Honest with Causal Language in Writing for Publication." *Journal of Advanced Nursing* 76 (6): 1285–88. https://doi.org/10.1111/jan.14311.

Tregoning, John S., Katie E. Flight, Sophie L. Higham, Ziyin Wang, and Benjamin F. Pierce. 2021. "Progress of the COVID-19 Vaccine Effort: Viruses, Vaccines and Variants Versus Efficacy, Effectiveness and Escape." *Nature Reviews Immunology* 21 (10): 626–36. https://doi.org/10.1038/s41577-021-00592-1.

VanderWeele, TJ. 2016. "Mediation Analysis: A Practitioner's Guide." *Annual Review of Public Health* 37. https://doi.org/10.1146/annurev-publhealth-032315-021402.

Waldner, David. 2015. "Process Tracing and Qualitative Causal Inference." *Security Studies* 24: 239–50.

Wieczorek, Aleksander, and Volker Roth. 2019. “Information Theoretic Causal Effect Quantification.” *Entropy* 21 (10): 975. https://doi.org/10.3390/e21100975.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yadav, Pragya D., Rima R. Sahay, Gajanan Sapkal, Dimpal Nyayanit, Anita M. Shete, Gururaj Deshpande, Deepak Y. Patil, et al. 2021. “Comparable Neutralization of SARS-CoV-2 Delta AY.1 and Delta in Individuals Sera Vaccinated with Bbv152.” https://doi.org/10.1101/2021.07.30.454511.

Yu, Bei, Yingya Li, and Jun Wang. 2019. “EMNLP-IJCNLP 2019.” In, 46644674. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1473.