Statistical Modeling: The Three Cultures

Adel Daoud\textsuperscript{1,2} Devdatt Dubhashi\textsuperscript{2}

\textsuperscript{1}Institute for Analytical Sociology, Linköping University, Linköping, Sweden,
\textsuperscript{2}Division for Data Science and Artificial Intelligence, Department of Computer Science and Engineering, University of Gothenburg and Chalmers University of Technology, Gothenburg, Sweden

Published on: Jan 26, 2023

DOI: https://doi.org/10.1162/99608f92.89f6fe66

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
ABSTRACT

Social scientists distinguish between predictive and causal research. While this distinction clarifies the aims of two research traditions, this clarity is blurred by the introduction of machine learning (ML) algorithms. Although ML excels in prediction, scholars are increasingly using ML not only for prediction but also for causation. While using ML for causation appears as a category mistake, this article shows that there is a third kind of research problem in which causal and predictive inference form an intricate synergy. This synergy arises from a specific type of statistical practice, guided by what we propose, the hybrid modeling culture (HMC). Navigating through a parallel debate in the statistical sciences, this article identifies key characteristics of HMC, thereby fueling the evolution of statistical cultures in the social sciences toward better practices—meaning increasingly reliable, valid, and replicable causal inference. A corollary of HMC is that the distinction between prediction and causation, taken to its limit, melts away.

Keywords: causal inference, prediction, social sciences, machine learning, artificial intelligence, data science.

Media Summary

There is a long-standing debate in the social sciences about the value of predictions and explanations. A prediction is usually about ‘using a procedure (a model) that takes some piece of information (data) and produces a statistical value (prediction) about the state of affairs.’ For example, to combat famines in a poor country, a researcher could set up an early-warning system for famine detection. This system could take a variety of daily input data, from weather input to the living conditions of the population in that poor country. Based on these inputs, the system could then produce daily predictions about the probability of famine. Although such a system would surely be a helpful tool in combating famines, it is less helpful in explaining why famines happen in the first place. Nonetheless, this sort of predictive logic is becoming increasingly common in social sciences, and beyond.

A causal explanation is a statement about ‘the difference between an outcome occurring when the event of interest is present versus absent.’ For example, if the event of interest is drought, and the outcome is famine, then a causal explanation must account for the difference in outcome when a drought is present versus absent. But it is physically impossible to observe both outcomes simultaneously for the same country. To handle this impossibility, social scientists tend to leverage a variety of research designs—from randomized control trials to natural experiments—to at least approximate their causal statements.

While predictive and explanatory sciences approach scientific inquiry radically differently, there is a new form of scientific practice that mixes both modes of thinking, combining the strength of the two modes. This practice we call the hybrid modeling culture (HMC). In this article, we discuss what HMC is, and how it supports scientific inquiry. We discuss how HMC relies on machine learning (ML)—a form of a procedure that
learns from data with little human guidance—which is usually used to produce predictive statements but is now used to produce explanatory statements. Besides causal inference, we show how HMC encourages the use of ML in other types of subproblems, from data acquisition to theory prediction.

Visit the web version of this article to view interactive content.

1. Introduction

Explaining social action; predicting social action. Traditionally, social scientists distinguish between predictive and causal research (Boudon, 2005; Elwert, 2013; Hedström & Ylikoski, 2010; Lundberg et al., 2021; Marini & Singer, 1988; Merton, 1968; Morgan & Winship, 2014; Risi et al., 2019; Shmueli, 2010; Watts, 2014). A *prediction* is a statement about the extent to which one or several events (input), when they occur, supply information about another occurring event (output). This association between inputs and output may or may not have a causal explanation—a mechanistic statement about why and how the input affects the output—attached to it. A *causal* statement is a counterfactual statement about a difference between an outcome occurring with an event (a policy, treatment, or exposure) activated versus deactivated (fully defined in Section 4.1). When this difference has a value other than zero, scholars take it as evidence that the exposure event causes the outcome.

While the distinction between predictive and causal statements has contributed to holding the truce among different quantitative research traditions (Freedman, 1991; Watts, 2014), unease is rising as scholars are increasingly using machine learning (ML) algorithms to analyze social phenomena (Bail, 2017; Lazer et al., 2020; Molina & Garip, 2019; Nelson, 2020; Shmueli, 2010; Turco & Zuckerman, 2017; Verhagen, 2022; Watts, 2017). ML is the study of how algorithms can learn from data (e.g., past social events) with no or little human guidance, thereby predicting new data instances (e.g., future social events) (Hastie et al., 2009). As ML excels in predicting social events (Bail, 2014; Boelaert & Ollion, 2018; DiMaggio et al., 2013; Kino et al., 2021; Molina & Garip, 2019; Mullainathan & Spiess, 2017; Salganik, Lundberg, et al., 2020), some scholars propose using ML algorithms for pure prediction problems “where causal inference is not central, or even necessary” (Kleinberg et al., 2015, p. 491).

Although many scholars agree that explaining social events requires causal statements, there is much less agreement on whether causal research also needs to be predictive (Boudon, 2005; Hedström & Ylikoski, 2010; Hofman et al., 2017; Keuschnigg et al., 2017; Marini & Singer, 1988; Merton, 1968; Shmueli, 2021; Verhagen, 2022). For example, Duncan Watts argues that “if [social scientists] want their explanations to be scientifically valid, they must evaluate them specifically on those grounds—in particular, by forcing them to make predictions” (2014, p. 313).
While we agree that there is a set of predominantly causal-research questions and likewise a set of predictive ones, our article shows that there is a third kind of research problem where causal and predictive statements form an intricate synergy. This synergy relies on further refinement of ML algorithms, and while a subset of this third kind of problem builds on Watts’s argument—what we call ML for theory prediction (discussed in Section 4.3)—there are at least two additional subsets relevant for social science research: ML for causal inference (Section 4.1) and ML for data acquisition (Section 4.2). To see how this third kind of problem is possible and relevant for social scientists, we must maneuver through a similar debate in the statistical sciences. This debate is cultural, in the sense that it forms a perspective that shapes the design and purpose of statistical models (including ML algorithms).

Two decades ago, statistician Leo Breiman (2001a) identified two cultures for statistical modeling. The data modeling culture (DMC) refers roughly to practices aiming to conduct model validation, and thus, statistical inference on one or several quantities of interest—distributions, model parameters, and alike. In the context of social sciences, such inferences often refer to defining a procedure $\hat{\beta}$ that estimates a true quantity $\beta$, to minimize the error $|\beta - \hat{\beta}|$ (Freedman, 1991). This true quantity $\beta$ is assumed to exist independently of the statistical model. For example, in any particular year and poverty definition (e.g., dollar a day), U.S. poverty levels are assumed to have a true level $\beta_{\text{poverty}}$, yet scholars can only approximate this level, $\hat{\beta}_{\text{poverty}}$, because they have to account for a variety of disturbances related to sampling and measurement. Similar disturbances exist in causal inference, which can be viewed as a particular type of statistical inference (Imbens & Rubin, 2015; Morgan & Winship, 2014; Pearl, 2009). For example, social scientists estimating a policy (causal) effect, $\hat{\beta}_{\text{policy}}$, of a new education program on school performance assume that this policy has a true effect, $\beta_{\text{policy}}$, yet this estimation is hampered by, among other things, the characteristics of students. A procedure is unbiased when the difference $\beta - E[\hat{\beta}_i]$ is zero (or negligible), over all possible realizations $i$.

Based on such statistical concepts, Breiman argued that DMC is the dominant mode of operation in statistics. While 20 years later, this culture has perhaps lost some of its dominant role in statistics because of the data science revolution, we observe that this culture is still the modus operandi—the leading practice of a group—in the social sciences and beyond (Goldthorpe, 2015). DMC is perhaps still the modus operandi because of the influence of the established scientific method for quantitative research in the social sciences, called the hypothetico-deductive scientific method (Hempel, 1965; Popper, 2002). This scientific method consists of cycles of deductively formulating a hypothesis from substantive theory, testing this hypothesis in a model and against data, and then, revising the theory based on empirical results (Costantini & Galavotti, 1986). Here, deductive has two meanings: first, from substantive theory, a scholar articulates a hypothesis about how two events $X$ and $Y$ are related; second, from this theory or by convention, a scholar stipulates a statistical model $f$ (often parametric and linear) for how this hypothesis of $X$ and $Y$ will be tested against data (Abbott, 1988).

While testing a model against data is an act of induction, the whole procedure follows a deductive process. The scholar seeks to model the generative process of the data and understand the process between $X$ and $Y$. 
Deductive reasoning uses universal propositions (hypotheses derived from general theories) to explain specific events. By *explanation*, we mean a theory that demonstrates how two or more events are mechanistically related (Goldthorpe, 2015; Hedström & Ylikoski, 2010). When enough evidence has been collected that challenges existing truths, an entire paradigm can fall in favor of a new one (Kuhn, 2012). The scientific method favors DMC over other modeling cultures because DMC supports causal reasoning.

The *algorithmic modeling culture* (AMC) refers to practices defining a procedure, $f$, that generates accurate predictions, $\hat{Y}$, about an event (outcome), $Y$ (Breiman, 2001a). By *accurate*, we mean predictions that are as similar as possible to the true event that $f$ has yet not encountered (Hastie et al., 2009). A *procedure* is an algorithm, or a function, that takes some input $X = x$, operates on this input $f(x)$, and then, produces an output $f(x) = \hat{Y}$. Often, this procedure is defined inductively (Costantini & Galavotti, 1986), that is, by letting the procedure learn from the patterns in the data, with little or no human guidance (Hastie et al., 2009). After learning the function from this *training data*, the procedure is used to make predictions on new data that was not present in the training data. The *test error* on this new data is a measure of how well the procedure *generalizes* to new data. While working in AMC, scholars often care about the statistical interpretability of the used procedure only so far that it furthers their pursuit to generate accurate predictions (Lipton, 2017). To identify causal relationships between $X$ and $Y$ is a peripheral question. This culture is the modus operandi of many strands in engineering, computer science, industry, and policy (Sanders 2019). As AMC procedures do not align with the hypothetico-deductive scientific method, this culture lacks subscribers from causally oriented social science research (Freedman, 1991; Molina & Garip, 2019).

The key difference between DMC and AMC is that the former is *process oriented* (i.e., modeling the generative process of the data), while the latter is *performance oriented* (i.e., building an emulator to match the predictive performance of a social system as closely as possible). In practice, this difference in orientation implies that DMC encourages using simpler (linear) interpretable models with few parameters, while AMC typically uses large complex (highly nonlinear) models for predictive accuracy disregarding interpretability (Rudin, 2019).

Despite the (seeming) incompatibilities of AMC with the hypothetico-deductive scientific method, among some research groups, AMC and DMC mix intensely. We argue that this mixing has formed a fertile spawning pool for a mutated culture: a *hybrid modeling culture* (HMC) where prediction and inference have fused into new procedures that reinforce one another. As Section 4 discusses, scholars use these procedures in the pursuit of explaining how two events are causally connected by blending $\hat{Y}$-prediction problems and $\hat{\beta}$-inference problems to the point that it is difficult to tell them apart (Athey & Imbens, 2019; Kino et al., 2021; Molina & Garip, 2019; Mullainathan & Spiess, 2017; Yarkoni & Westfall, 2017). One such procedure, which we discuss in Section 4.1, is the use of machine learning in the service of causal inference (Künzel et al., 2019). A fused procedure is still compatible with the hypothetico-deductive scientific method but stretches beyond it because it allows for a much larger portion of inductive reasoning (Nelson, 2020). Such reasoning infers generalized...
claims from particular observations. While this hybrid culture does not occupy the default mode of social science practices, we argue that it offers an intriguing novel path for applied social sciences.

This article aims to identify key characteristics of what we named HMC, thereby facilitating the scientific endeavor and fueling the evolution of statistical cultures in the social sciences toward better practices. By better, we mean increasing valid, reliable, and reproducible practices in analyzing causal relationships (Lundberg et al., 2021; Morgan & Winship, 2014; Watts, 2014). By valid and reliable, we mean practices that lead to studies that ‘systematically measure what they intend to measure,’ and ‘consistently generalize to the target population of interest,’ respectively. Increasing reproducible refers to practices that make studies less dependent on the architecture of deductive models than what is currently the case in DMC (e.g., linear models). Even if there are several reasons for the reproducibility crisis in the sciences (Camerer et al., 2018), the use of less flexible models will inadvertently squeeze complex social data into an unsuitable format. That unsuitability adversely affects the robustness of applied research.

We execute our account relying on argument by example, meaning that we will selectively review trends in applied social science research as proof of the existence of HMC. From these examples, we will pinpoint the defining characteristics of HMC.

Before discussing HMC, we suggest two trends that nourish its emergence. First, to a considerable extent, HMC has emerged from applied computational research that closely interfaces with statistics and computer science—that is, data science (Efron & Hastie, 2016). This close interfacing in the social sciences is known as computational social science, which denotes any scientific study that develops or uses computational methods to typically large-scale and complex social and behavioral data (Keuschnigg et al., 2017; Lazer et al., 2020). The stream of new data sources—administrative data, social media, digitalized corpora, and satellite images—explain the relevance of such a computational approach (Jordan & Mitchell, 2015). Similar computational approaches exist under the brands of digital humanities (Gold, 2012), computational psychology (Sun, 2008), computational economics (Tesfatsion & Judd, 2006), computational epidemiology (Marathe & Vullikanti, 2013; Salathé et al., 2012), and computational biology (Noble, 2002), to mention a few. All these computational approaches emerged by the beginning of the 21st century, and they provide clues to why DMC or AMC are insufficient alone to cover the new demands of the scientific endeavor: that is, to provide systematic explanations of events of reality, and thereby to deepen our knowledge of them (Bhaskar, 2008; Hedström & Ylikoski, 2010; Watts, 2014).

A second feeding ground of the evolution of HMC is the causal-inference revolution. While a randomized control trial (RCT) remains the safest way to rinse out the contaminating effect of confounding and the least assumption-demanding method to identify causality (Fisher, 1935; Morgan & Winship, 2014), scholars face ethical and practical limitations when applying an RCT to social settings (Deaton & Cartwright, 2016). For example, to estimate the causal impact of events such as economic crises (Daoud et al., 2017; Elder, 1998), famines (Sen, 1981), or climate change (Arcaya et al., 2020; Hsiang, 2016) on children’s well-being, scholars
would need to administer such events to a treatment and control group of children. Ethical and practical limitations of RCT are two sources fueling the causal revolution that has resulted in a myriad of new approaches tailored to inferring causality from observational data (Angrist & Pischke, 2014; Hedström & Manzo, 2015; Hernan & Robins, 2020; Imai, 2018; Imbens & Rubin, 2015; King, 1998; Morgan & Winship, 2014; Pearl & Mackenzie, 2018; Peters et al., 2017; van der Laan & Rose, 2011; VanderWeele, 2015). As this revolution has partly evolved from computer science, partly from statistics and economics, scholars have creatively combined tools from DMC and AMC. As HMC synthesizes the strengths of DMC and AMC, this synthesis has resulted in practices better adapted to 21st-century social science requirements than what DMC or AMC alone can offer.

The remainder of our argument is less concerned with why HMC has emerged and more with characterizing it.

### 2. Thinking Predictively and Inferentially

#### 2.1. The Data Modeling Culture (DMC)

Scholars in quantitative social sciences think and operate mainly through the hypothetico-deductive scientific method (Danermark et al., 2002; Kuhn, 2012; Watts, 2014). This method is a philosophy of science that defines how scientific inquiry should be conducted (Hempel, 1965; Popper, 2002). Using substantive theories, scholars articulate their causal and descriptive knowledge in falsifiable hypotheses and operationalize them in data. Thus, this method is ‘hypothetico,’ implying that hypotheses are postulated by specifying a statistical model representing how the data (events) are generated. These models represent scholars’ best guess of the data-generating process of the phenomena studies (elections, poverty, inequality, criminality, etc.). As statistical models are postulated, the scientific method is deductive, because this method aims to test whether the assumed data-generating process of the model matches the sampled data (King, 1998). Based on that model and sampled data, scholars evaluate how likely the null hypothesis—for example, a statistic that two events $A$ and $B$ are unrelated—is receiving support in data. When the null is receiving support, the sample statistic coincides with the postulated model. Intuitively, that is what a high $p$-value means: a probabilistic claim of how likely we are to encounter the postulated statistic under the null. We are willing to reject the null hypothesis if the sample value we encountered is unlikely to occur, and we favor the alternative hypothesis—that $A$ and $B$ are related.

Despite suffering from paradoxes (e.g., the raven paradox, irrelevant conjunctions) (Huber, 2022; Schultz, 2018) and other problems (e.g., the problem of underdetermination) (Crupi & Tentori, 2010), the hypothetico-deductive method offers quantitative-applied research a way to reason about statistical inference. Although we remain agnostic about whether scholars should rely on the hypothetico-deductive method for conducting inference, scholars require some comparable framework for producing uncertainty estimates for inference. Because scholars revise their theories (knowledge) based on that inference, they sample more data, and refine
their statistical models. And so, the circle of science continues. The requirement of testing substantive theories through an interpretable statistical model is one of the appeals and endorsement of DMC (Breiman, 2001a).

Throughout cycles of knowledge production, deduction and induction in statistical analysis form complementary scientific phases, yet they represent two fundamentally different ways of producing social-scientific knowledge based on statistical data. While Carnap (1962) is one of the leading proponents of inductivism, Popper (2002) defends deductivism. The opposition is rooted in the difference between confirmation (induction) and falsification (deduction). In the language of statistics, confirmation is about methods of estimation and falsification is about tests of significance. These procedures—estimating and testing—operate in different stages in knowledge production: the stage of estimation and stage of testing, yet jointly capture vital aspects of the scientific endeavor.

As previously mentioned, because of DMC’s deductive flavor, its modeling culture relies mainly on the context of falsification; conversely, AMC’s inductive preferences lead to a statistical culture favoring the context of estimation. While AMC is concerned with prediction, it relies on estimation for prediction and not the interpretation of model parameters.

To compare DMC and AMC—and characterize HMC—we define the following terminology. Scholars formulate, test, and develop social theories, $T_1, \ldots, T_k$, about a causal (social) system. A causal system is a set of events and relationships between events in a domain of interest. Because elements of the causal system rarely reveal themselves directly to the human senses, scholars theorize about the existence of events and their causal relationships (Bhaskar, 2008; Hedström, 2005). By theory, $T_k$, we mean a set of concepts that enable formulating descriptions, predictions, hypotheses, or explanations about events populating a causal system (Swedberg, 2017). Each $T_k$ maps into a directed acyclic graph (DAG), $G_k$, that formalizes and visualizes a potential manifestation of the causal system of interest (Pearl, 2009). While two or more theories often compete for proposing the best explanation—meaning how well they account for the mechanisms generating observed data—they do not have to be mutually exclusive. Scholars can formulate theories at various abstraction levels, but to test them empirically, they need to match what is measurable. Thus, there is a social-scientific preference in quantitative research to engage with middle-range theories rather than grand theories (Merton, 1968).

Poverty and famine research serves as an illustration. Figure 1 shows the progression of knowledge under DMC with three stylized DAGs, $G_1$, $G_2$, and $G_3$ competing to explain famines. A debate raged between Malthusians and Senians on whether food scarcity is a necessary event to cause famines (Devereux, 2007; Sen, 1981)—a debate that still influences human ecology, sustainability, and adjacent research (Daoud, 2018). Three centuries ago, Thomas Malthus argued that while population size increases geometrically, food supply increases arithmetically (Malthus, 1826). Because population size will outstrip food supply, famines will eventually emerge to balance their relationship. If we assume linear relationships among these mechanisms, the effect of scarcity on the probability of famine arising can be captured by a parameter $\beta_1$; the effect of food
supply on the probability of scarcity can be encoded by a parameter $\beta_2$; and the effect of population size on the probability of scarcity can be quantified by a parameter $\beta_3$. Based on Malthus’s theory $T_1$ and these parameters, we can specify $G_1$ that explains why a famine arises, as shown in Figure 1. Amartya Sen challenged this explanation by showing that famines—at least in the modern era—can arise even when there is sufficient or abundant food (Sen, 1981). Especially when social inequality is high, vulnerable groups run a higher risk of unemployment than other groups. Unemployment causes a loss of income and individuals’ capability to purchase food. This loss of capability—what Sen named entitlement failure—results in starvation, which is defined by a parameter $\beta_4$. In Sen’s theory $T_2$, the causal system of $G_2$ is a better explanation of how famines arise. Although Sen acknowledged that population size and food supply shortage can cause famines, such shortage is “one of many possible causes” during the last century (Sen, 1981, p. 1). Thus, Sen’s theory argues that population size and food supply have roughly no effect on the probability of famines, that is, $\beta_3, \beta_2 \approx 0$. 


Subsequent theoretical development uses both theories to offer an even more robust approach to explaining events of famines (Daoud, 2017). A critical conceptual movement is to disentangle societal and individual-level starvation, yielding additional parameters. As $G_3$ shows, while Malthus’s theory explains when famines are likely to arise at the societal level, it cannot explain which individuals might starve to death. Sen’s theory

---

**Figure 1.** Directed acyclic graphs depicting stylized causal systems of famine. This figure represents three stylized causal systems of how famines are caused. The $\beta$ parameters on the arrows represent the average causal effect between nodes.
identifies these individuals by their failing food entitlements (Reddy & Daoud, 2020). Additionally, the entitlements-causal path ($\beta_4$ via $\beta_6$ to $\beta_7$) shows that famines can arise even if there is no societal scarcity.

As exemplified in the research progress of poverty and famine research, DMC practices encourage social scientists to pursue ever deeper knowledge production. Scholars aim to quantify, describe, and evaluate key events and their relationships. A *description* is a statistical inference about aspects of the distribution of one or more random variables, represented as a node in a DAG. A *random variable* encodes a distribution of the event of interest. For example, let $FAMINE$ be a binary random variable measuring the probability that an event of a famine occurs in a well-defined time (e.g., in the years 1900 to 2000) and space (e.g., Europe, Asia, or Africa). Then, the expectation $E[FAMINE]$ captures the average occurrence of famines in that time and space. In an applied setting, this average is estimated by the empirical average, that is, the number of famines that occurred divided by the maximum number of famines that could have occurred. As $FAMINE$ is a binary random, the quantity $E[FAMINE]$ equals the probability of $P(FAMINE)$. Small caps denote variables in a DAG. A description of two or more random variables refers to associations between several events. *Associations* (correlation) may or may not be causal—specific assumptions need to hold for identifying causal associations (Hernan & Robins, 2020; Imbens & Rubin, 2015; Pearl, 2009).

When randomization of treatment (or exposure) is infeasible, scholars have to rely on causal assumptions about observational data (Rosenbaum, 2020). Besides methodological issues, scientific debates often refer to the content of a causal system supplied by such data. For example, scholars debate what variables populate such a system and how they affect each other. This content also defines the conditions under which empirical associations may be interpreted as causal or confounded. As Section 4.1 discusses, a confounded association is an association between two variables $W$ and $Y$ that is partly determined by a common cause, $C$. If a supposedly causal association is confounded, then that association between $W$ and $Y$ is biased (Hernan & Robins, 2020). A *causal association* is the portion of all association that arises entirely to one variable (e.g., $W$) affecting the values of another variable (e.g., $Y$). To statistically capture causal associations, scientists analyze the joint distribution of the variables of interest, that is $p(Y, W, C)$, how this joint distribution factorizes, and their underlying structural causal models.

A factorization defines the causal order among events. For example, the joint distribution $p(FAMINE, SCARCITY, POPULATION, FOODSUPPLY, ENTITLEMENTS)$ has many potential factorizations depending on substantive theory. While Malthus’s theory, as defined in $G_1$, stipulates the following factorization of how famines arise,

$$p(FAMINE, SCARCITY, POPULATION, FOODSUPPLY) = p(FAMINE|SCARCITY) p(SCARCITY|POPULATION, FOODSUPPLY) p(FOODSUPPLY) p(POPULATION),$$

Sen’s (1981) $G_2$ declares that the following factorization is the best approximation of how famines arise,
T

hese factorizations follow the order of nodes and arrows in Figure 1. The scientific and policy differences are large, depending on whether $G_1$ or $G_2$ is the best representation of reality (Daoud, 2017). If $G_1$ best represents the causes and effects of famines, then policymakers should produce more food or contain population growth to counter famines; conversely, if $G_2$ is a better representation, then policymakers should follow this theory stipulating that they ought to reduce social inequality for reducing the probability of famines. Such mixing of causal reasoning and ethics is an emerging field in computer science (Daoud, Herlitz, & Subramanian, 2022; Kusner et al., 2017).

While a factorization defines the conditional dependencies in a causal system, a structural causal model (SCM) moves one step further in specificity by defining the direction of causality. That direction is defined by encoding the functional relationship among all the variables in that system. A functional relationship is a model $f$ that specifies a one-directional mapping between an outcome (effect) and input (causes) variables. That model can be of any functional class, parametric or nonparametric, and reflects how a theory quantifies a causal system. For example, based on Malthus’s theory, as defined in $G_1$, we can define the following SCM, where the models $f_k$ are nonparametric with noise terms $e_k$, where the index is over a step in the SCM,

\[
p(\text{FAMINE, SCARCITY, POPULATION, FOODSUPPLY, ENTITLEMENTS}) =
p(\text{FAMINE|SCARCITY}) \cdot p(\text{SCARCITY|ENTITLEMENTS}) \cdot p(\text{ENTITLEMENTS}) \cdot p(\text{FOODSUPPLY}) \cdot p(\text{POPULATION}),
\]

In DMC, to estimate the causes and effects of famines, scholars collect famine data and test their stipulated models. The model that fits the sample best receives scientific support. Scientific debates tend to amplify when different samples yield support for different models. Scholars evaluate support for a statistical model by interpreting how different factorizations match the sample. While scholars can use many different models, DMC-influenced scholars often use linear models to retain interpretability (Lipton, 2017). A linear model of $G_1$ could then have the following stylized statistical form, $\text{FAMINE}= f_1(\text{SCARCITY}, e_1)$, where $f_1(\text{SCARCITY}, e_1) = c_0 + \beta_1 \text{SCARCITY} + e_1$. In this example, a hypothesis operationalizes an aspect of $T_1$, for example, that $\beta_1$ is different from zero, and DMC-operating scholars use that linear model to test this hypothesis. Generally, a hypothesis operationalizes an aspect of $T_k$, stipulating the existence of a causal relationship (edges) between random variables (nodes). Scholars use interpretable statistical models so that they can imprint their hypothesis into these models. While linear models tend to simplify social reality too much (Abbott, 1988), linear models are popular because they make this imprinting straightforward (Rudin, 2019).
The goal of testing social theories through interpretable statistical models explains why a DMC-operating scholar shies away from a kitchen-sink method—that is, one that throws all the predictors one can find into an algorithm, and then, letting the algorithm regress the outcome of interest on them. As previously defined, ML is the subdiscipline of computer science that studies how algorithms can learn from data (Efron & Hastie, 2016; Hastie et al., 2009). As many ML algorithms are nonparametric, it is unclear how to unpack and interpret them in a direct manner, as is commonly done for linear (parametric) models (Lipton, 2017). Without a clear theoretical rationale for using ML, DMC scholars see little value in using them in the scientific process (Breiman, 2001a).

2.2. The Algorithmic Modeling Culture (AMC)

Machine learning algorithms lie at the heart of many AMC practices. A key assumption of AMC is that a system produces an association between a given set of inputs, $\mathbf{X}$, and a particular output, $\mathbf{Y}$. The relationship between $\mathbf{X}$ and $\mathbf{Y}$ may or may not be causal. The overarching goal is to develop a model $f$ that operates on these inputs, producing the best possible predictions $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ that $f$ has not been observed yet. For our famine example, $f$ can be any ML algorithm that performs well in predicting famines, a critical component of famine-early-warning systems. The Famine Action Mechanism—a collaboration among the World Bank, United Nations, Food and Agricultural Organizations, and others—and the Famine Early-Warning-System-Network (led by the United States Agency for International Development) are examples tailored to minimize the outbreak of famines. Here, the primary focus is to predict famine and not explain why famine happens, contrasting with DMC famine models.

The relationships between $\mathbf{X}$ and $\mathbf{Y}$ are not necessarily causal in AMC. Figure 2 shows a stylized graphical representation of the system of associations among inputs and outputs. This graph does not represent a causal relationship because some nodes are undirected (those between all $\mathbf{X}$s), and thus, we call it a Bayesian network (Pearl, 2009). All DAGs are Bayesian networks, but not all Bayesian networks are DAGs. In a Bayesian network, an edge denotes an association between two nodes with a noncausal interpretation. In Figure 2, all inputs $\mathbf{X}_k$ are connected because the association is assumed to flow in all directions. If we were to use a linear model to represent the fully connected graph, even then would we end up with many parameters to estimate, thereby hampering interpretability: first, there are $\beta_1, \ldots, \beta_k$ association terms between $\mathbf{Y}$ and each $\mathbf{X}_k$; second, there are $\beta_{k+1}, \ldots, \beta_{k+\left(\frac{k}{2}\right)}$ pairs of correlation parameters among all the inputs $\mathbf{X}_1, \ldots, \mathbf{X}_k$. Because scholars evaluate the predictive performance of $f$ comparing $\hat{\mathbf{Y}}$ using a held-out set $\mathbf{Y}$, scholars pursue to interpret these association only as a subordinate priority, if at all (Doshi-Velez & Kim, 2017).
Nonetheless, explainable AI is a recent area in machine learning that develops techniques for probing the inner workings of ML (Doshi-Velez & Kim, 2017; Miller, 2019; Molnar et al., 2020). These techniques attempt to

Figure 2. Bayesian network in the algorithmic modeling culture. The figure represents a covariate set associated with an outcome, without any causal assumptions.
enrich a model’s prediction with an explanation for why a model made a certain prediction. Especially suitable for the deep learning models, as they tend to be the most complex algorithms with millions of parameters, one trains an ML (black-box) model and produces a set of predictions. For each prediction, one probes what parts of the parameters space were activated, and thereby creates an explanation or interpretation of the ML model’s behavior. For example, to let an image-recognition algorithm predict whether a neighborhood is wealthy from satellite images, scholars can probe the algorithm and identify what part of the image made the algorithm react in a certain way. Was it the number of backyard swimming pools, the absence of expensive cars, or some other characteristics that led the algorithm to predict the wealth of neighborhoods?

Interpretable artificial intelligence (AI) seeks to clarify the decisions suggested by predictive models by moving one step beyond explainable AI (Rudin, 2019). While explainable AI focuses on trying to explain the inner mechanics of black-box models (e.g., through activation maps in image-processing models), interpretable AI focuses on creating models that produce interpretable results in the first place. These models are critical in high-stakes decision-making situations, such as when and who should be prioritized for health care or detained or bailed in criminal justice or which neighborhoods should be selected for public policy interventions.

However, under the spell of AMC, the goal of explainable or interpretable AI is seldom set to identifying cause and effect. As the main goal of AMC is not to develop causal knowledge, $T_k$, about a system, AMC demotes causal reasoning (Pearl & Mackenzie, 2018). Despite the lack of causal reasoning, AMC exhibits many ML innovations across several domains. In robotics, autonomous vehicles are capable of driving far distances without human supervision; in the arts, the Next Rembrandt project has shown how image-recognition algorithms can recreate a Rembrandt painting at the level of mastery that humans have difficulty telling them apart from an original Rembrandt; similarly, in music, MuseNet composes symphonies at par with Mozart, Chopin, or Beethoven; in linguistics, OpenAI’s Generative Pre-trained Transformer 3 (GPT-3) writes articles, songs, and manuscripts with impressive coherency; in gaming, DeepMind’s algorithms AlphaStar and AlphaGo compete at the grandmaster level in the computer game Starcraft II (Vinyals et al., 2019) and the board game Go (Silver et al., 2017), respectively.

Although much remains to be proven before the same algorithm—strong artificial intelligence—can roam across all these domains, AMC innovations are noteworthy because scholars have developed each algorithm without explicitly knowing the causal connections among events (Domingos, 2015). Machine learning models learned the relevant associations from data, with little supervision.

The advancements of AMC resonate with Karl Pearson’s idea that statistical correlation—predictability—between $X$ and $Y$ is what scholars should search for to advance science. For Pearson, correlation is causation. He argued,

> Take any two measurable classes of things in the universe of perceptions, physical, organic, social or economic, and it is such a dot or scatter diagram, which we reach with extended observations. In some
cases the dots are scattered all over the paper, there is no association of A and B; in other cases there is a broad belt, there is only moderate relationship; then the dots narrow down to a “comet’s tail,” and we have close association. Yet the whole series of diagrams is continuous; nowhere can you draw a distinction and say here correlation ceases and causation begins. Causation is solely the conceptual limit to correlation, when the band gets so attenuated, that it looks like a curve. (Pearson, 1911, p. 170)

Although scholars have produced noteworthy innovations in AI using the principles of AMC, we argue that the absence of causal reasoning is a major limitation for the advancement of scientific knowledge for applied social sciences and related scientific domains (Darwiche, 2017; Pearl & Mackenzie, 2018). Developing ML that paints like Rembrandt and composes symphonies like Mozart, but not knowing exactly what causes what in their respective machinery, is insufficient for the advancement of social-scientific knowledge. Similarly, training an ML-powered early-warning system for famine detection will support policymakers in reacting to starvation, but not remedying starvation. That is, if scholars are unable to unpack what affects what in a famine system, then the science of famines gains little from hindering famines to arise in the first place.

Even if DMC suits the scientific endeavor better, it suffers from at least two limitations. First, because of the suspicion toward AMC-style predictions, DMC scholars tend to rely on analog methods to collect information about the causal systems (Salganik, 2017). Surveys, experiments, and interviews are examples of analog methods as they give full (human) control over the data collection process. However, in the digital age, these analog approaches limit the speed and type of data collected that could be processed and used for the scientific endeavor. Second, in agreement with Breiman (2001a), we argue that a more severe limitation of DMC is that it relies mainly on model validation. A scholar formulates a statistical model and then tests that model against data. Using various goodness-of-fit metrics, this scholar then draws a set of conclusions. Yet, these conclusions are often more telling about the assumed model’s structure and less about the causal system of interest. If the statistical model is a poor representation of this causal system—for example, that the relation between $X$ and $Y$ is not linear—these conclusions may be misleading, or nonreproducible (Breiman, 2001a).

Table 1 summarizes the key characteristics of DMC and AMC. As written in the first row, although famines exemplify these characteristics, changing food supply to $X$ and famines to $Y$, one can generalize these characteristics to any scientific domain of interest. The goal of DMC is to identify the causal quantity $\hat{\beta}$ between in how food supply affects famines. In AMC, the goal is shifted from inferring causality to predicting famines, $\hat{Y}$, using any relevant input data, $X$. While the former stipulates untestable statistical assumptions, the latter relies on black-box models. All these respective limitations notwithstanding, different disciplines have made breakthroughs under the cultural influence of AMC and DMC. A warranted question is then: to what extent can they be synthesized to further the social-scientific endeavor?
Table 1. Central practices of two statistical cultures.

| Exemplifying question | Data modeling culture (DMC)                                                                 | Algorithmic modeling culture (AMC)                                                                 |
|-----------------------|------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Goal                  | Estimating unbiased parameters for causal estimation, to populate the magnitudes of the edges of a directed acyclic graph (DAG). | To develop and train an algorithm $f$ for accurate prediction.                                      |
| A key assumption      | Assuming a DAG, a stipulated and interpretable statistical model such as $y_i = c_0 + \beta w_i + e_i$ produces unbiased estimates of the true causal quantity $\beta$. | The algorithm $f$ can produce accurate predictions of $Y$ from data source, $D$.                   |
| Limitation            | Although the parametric model is interpretable, its statistical structure may be a poor representation of the causal system. | Although $f$ produces accurate predictions, the model is a black-box restricting causal interpretations. |
| Quantity of interest  | $\beta$                                                                                  | $\mathcal{Y}$                                                                                   |

Note. a) in the equation $y_i = c_0 + \beta w_i + e_i$, the outcome is $y_i$ and the treatment is $w_i$. The variable $c_0$ is the intercept and $e_i$ is the residual.

3. A Hybrid Statistical-Modeling Culture: A Unifying Framework Fueling the Evolution of Scientific Practices

By upholding disciplinary traditions, university departments also inadvertently create cultural silos where DMC and AMC practices dwell (Lazer et al., 2020). Changing these traditions is challenging. Nonetheless, through focused efforts (e.g., the Summer Institutes in Computational Social Science founded by Chris Bail and Matthew Salganik) or forming interdisciplinary centers (e.g., Harvard’s Institute for Quantitative Social Science, the Santa Fe Institute, or the Institute for Analytical Sociology, Sweden) where the isolating effect of departmental silos are being dismantled, a new hybrid modeling (HMC) culture is emerging.

As HMC is an evolutionary descendant of DMC and AMC, it benefits by copying useful elements from each of the two-parent cultures and mutating them into new practices. Following DMC’s scientific goal, HMC still submits to the overarching aim of science: to identify and explain the causal link between any events $X$ and $Y$ of interest (Pearl & Mackenzie, 2018). This aim manifests in the advancement of substantive theories, explanations, and hypotheses that cyclically gets tested with statistical models and against data. Nonetheless, instead of relying on the commonly used statistical models in DMC, HMC leverages the arsenal of ML algorithms developed under AMC, thereby increasing scholars’ modeling power (Kuang et al., 2020; Künzel et
al., 2018; Peters et al., 2017; van der Laan & Rose, 2011, 2018). In combining inference and prediction, the result of HMC is that the distinction between $\hat{Y}$ and $\hat{\beta}$—taken to its limit—melts away.

We discuss our melting-away argument by describing three HMC practices, where each practice captures an aspect of the scientific cycle. Table 2 shows an overview of what these three practices constitute: ML for causal inference, ML for data acquisition, and ML for theory prediction. Although these three goals exist partly in DMC (i.e., parametric inference) and AMC (i.e., ML-style prediction), HMC fulfills them by blending inferential and predictive thinking. The three HMC practices that we will discuss in the next section combine ML prediction and inference to such a high degree that neither DMC nor AMC can comfortably host them. For example, while HMC inherits the goal of causal inference from DMC, there is no DMC equivalent for letting an algorithm discover how a DAG should be specified from data alone, with little or no guidance from substantive theory. Training such algorithms is an HMC-specific problem, called causal discovery (Glymour et al., 2019; Peters et al., 2017). As discussed in the next section, under a set of assumptions, causal-discovery algorithms can recreate a social system by suggesting potential DAGs from observational data. These algorithms are likely useful when social theory is weak, and therefore, scholars aim to generate hypotheses inductively.

Table 2. Central practices of the hybrid modeling culture (HMC).

| Exemplifying question | Machine learning (ML) for causal inference | ML for data acquisition | ML for theory prediction |
|-----------------------|-------------------------------------------|------------------------|-------------------------|
| Goal                  | What is the causal relationship between food supply and famines? | Can food availability be measured from satellite images? | How well does the Malthusian theory of famines predict new famines? How does it compare to a Senian theory? |
| A key assumption       | Imputing potential outcomes for causal estimation, to populate the magnitudes of the edges of a directed acyclic graph (DAG). | Producing new indicators from digital sources, $D$, to populate the nodes of a DAG. | Comparing the predictive power of two or more theories’ DAGs, $\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2, ..., \hat{\mathcal{G}}_k$, for new realizations of an outcome, $Y$. |
| Quantity of interest   | $\hat{\tau} = \hat{\beta}^1 - \hat{\beta}^0$ | $\hat{X}, \hat{W}, \hat{Y}$ | $\epsilon_{\mathcal{G}_k} \approx Y - \hat{Y}_{\mathcal{G}_k}$ or $\epsilon_{\mathcal{G}_k} \approx \tau - \hat{\tau}_{\mathcal{G}_k}$ |

Note. The Exemplifying question row merely highlights one research question in a set of myriad questions that can be formulated

### 3.1. ML for Causal Inference

Before describing how ML aids in inferring causality in HMC (column one in Table 2), we will refine our definition of what we mean by causal inference. Table 3 illustrates an observed data matrix of four individuals with fictitious variable values.
Table 3. A toy data set illustrating the fundamental problem of causal inference.

| Unit | $Y$ | $Y^1$ | $Y^0$ | $W$ | $\tau$ | $X$ |
|------|-----|-------|-------|-----|-------|-----|
| Jane | 20  | 20    | ?     | 1   | ?     | 10  |
| John | 30  | 30    | ?     | 1   | ?     | 11  |
| Joe  | 25  | ?     | 25    | 0   | ?     | 10  |
| Jan  | 22  | ?     | 22    | 0   | ?     | 11  |

Note. All the numbers provided in all the cells are fictitious. They are generated to exemplify how a dataset could look like, and that there will always be unobserved potential outcomes, signified by the question marks.

We define the causal effect of a (binary) variable, $W$, on $Y$ in terms of potential outcomes. Instead of merely recording each individual’s outcome as observed by the data, $Y_i$, we assume that each individual $i$ has two potential outcomes (Imbens & Rubin, 2015). One potential outcome represents the outcome when the individual takes the treatment (that is, $W_i = 1$), $Y^1_i$, and one where they do not take it, $Y^0_i$. The causal effect, $\tau_i$, for each individual $i$ is then the difference between these two potential outcomes:

$$\tau_i = Y^1_i - Y^0_i.$$

If we could observe both potential outcomes, we could then directly compute $\tau_i$ and thus identify individual-level causal effects. However, the observed outcome—as supplied by the data—is a function of both the treatment and the two potential outcomes, $Y_i = (1 - W_i) Y^0_i + W_i Y^1_i$. This function shows that the observed data, exemplified in Table 3, reveals only one of these two potential outcomes, yet both are required to identify causal effects. This impossibility of observing both potential outcomes is known as the fundamental problem of causal inference. Much of the causal-method development pertains to reasoning about identifiability and defining procedures for calculating causal effects from observational data (Hernan & Robins, 2020; Imbens & Rubin, 2015; Pearl, 2009; Peters et al., 2017). Identifiability means articulating a set of assumptions that allow a model to calculate a causal effect from observed data.

Vibrant literature in the intersection of computer science, econometrics, and statistics combine ML and causal methodology to identify causal effects (Athey et al., 2019; Athey & Imbens, 2017; Chernozhukov et al., 2018; Hedström & Manzo, 2015; Hernan & Robins, 2020; Hill, 2011; Hirshberg & Zubizarreta, 2017; Imai, 2018; Kino et al., 2021; Pearl & Mackenzie, 2018; Peters et al., 2017; Sekhon, 2009; van der Laan & Rose, 2011; VanderWeele, 2015). A reoccurring theme in this intersection is the many creative combinations of methods...
where predictive AMC-type algorithms are used in DMC-type causal inference. These causal methods summarize into at least four types of method combinations.

### 3.1.1. ML imputes potential outcomes

In the first combination, scholars use ML to impute (predict) potential outcomes (Daoud & Dubhashi, 2021). As observed data only reveal half of the potential outcomes, they regard the other half of the data as a missing data problem (Imbens & Rubin, 2015). One way of handling this fundamental problem is to identify conditions for imputing these data to populate all the $Y_i^1$ and $Y_i^0$ cells, based on the similarity of covariates $X$. These imputation procedures rely on common identifiability assumptions. One such central assumption is *conditional independence* (also known as *conditional ignorability* and *conditional exchangability*), $Y_i^1, Y_i^0 \perp W | X$. This mathematical statement means that the treatment is as-if randomly assigned conditional on one or more covariates.

Because ML excels in prediction tasks compared to commonly used parametric models, HMC-influenced scholars have developed many different procedures to predict (impute) potential outcomes (Künzel et al., 2018). For example, the *T-learner*—‘T’ stands for ‘two’—procedure defines one ML-algorithm $f_{w=1}(x_i) = E[Y = y_i | W = 1, X = x_i]$ trained on the treated group and another algorithm $f_{w=0}(x_i) = E[Y = y_i | W = 0, X = x_i]$ trained on the control group. Depending on the scientific problem, the scholar defines the type of algorithm—a Lasso, neural network, a random forest, or a collection of algorithms (an ensemble). After training, $f_{w=1}$ imputes potential outcomes for the control group and $f_{w=0}$ imputes these outcomes for the treated group. Based on the toy data in Table 3, $f_{w=1}$ trains on the variables of Jane and John, and imputes $Y_i^1$ for Joe and Jan; likewise, $f_{w=0}$ trains on Joe and Jan, and imputes $Y_i^0$ for Jane and John. Then, each individual-level effect is obtained by taking the difference $\hat{\tau}_{i \text{treated}} = Y_i^1 - \hat{Y}_i^0$ for the treated group and $\hat{\tau}_{i \text{control}} = \hat{Y}_i^1 - Y_i^0$ for the control group. To calculate the average causal effect, this procedure culminates by taking the weighted average over these individual-level effects in the treated and control groups, respectively, $\hat{\tau} = \pi \frac{1}{n_1} \sum \hat{\tau}_{i \text{treated}} + (1 - \pi) \frac{1}{n_0} \sum \hat{\tau}_{i \text{control}}$, where $\pi$ is the portion of treated individuals.

The T-learner algorithm is one of several causal-estimation methods, but common to most of these ML algorithms is the procedure of imputing potential outcomes (Künzel et al., 2018) or imputing the treatment effect directly (Athey et al., 2019; Nie & Wager, 2017). Consequently, in these sorts of HMC practices, $\hat{\tau}$-problems of HMC subsume $\hat{\beta}$-problems of DMC.

Imputing potential outcomes enables flexible modeling. As ML algorithms are flexible, they can automatically find the best functional form $f$ instead of relying on a scholar to select a model. As discussed in Section 3, a scholar under the influence of DMC would most likely articulate a set of assumptions for when the causal effect is identified (e.g., in a DAG) and specify a linear model to estimate this effect. For example, in our famine example depicted in Figure 1, a Malthusian scholar would argue that SCARCITY$_i$ (of food) causes...
This scholar may show the appropriateness of their assumption in the DAG, $G_1$, and proceed to specify the following stylized statistical model, $\text{FAMINES}_i = c_0 + \beta_1 \text{SCARCITY}_i + e_{1i}$, where $i$ indexes famine events. This model imprints the causal effect in the parameter $\beta_1$. If the true relationship between famines and scarcity is linear, this statistical model will capture the desired causal effect by extrapolating between famine cases where scarcity was observed and where it was not. However, in most scientific domains, a scholar’s preference to use a linear model follows more the desire to readily and transparently interpret a statistical model rather than following this scholar’s knowledge that the complexities of reality are truly linear (Abbott, 1988; Lipton, 2017). Although linear models can capture nonlinearities via a variety of transformations, to model a nonlinear reality, ML for causal inference offers a more robust alternative by approximating the best functional form (van der Laan & Rose, 2018).

If a famine scholar followed the statistical practices of HMC instead of DMC, this scholar would formulate the same causal goal (estimand) in the shape of a $\mathcal{T}$-problems, and now use an ML algorithm to estimate the causal effect by imputing potential outcomes (Künzel et al., 2018; Lundberg et al., 2021). Using the same DAG, $G_1$, the scholar would, for example, use a T-learner to impute the probability of famine in cases where scarcity is present and where scarcity is absent. Although both $\hat{\tau}$ and $\hat{\beta}$ quantify the same estimated causal effect, their key difference resides in that $\hat{\beta}$ often refers mainly to a parametric modeling setting, whereas $\hat{\tau}$ refers to any (nonparametric) setting—a statistical-modeling nomenclature. To calculate $\hat{\tau}$, the scholar benefits from the algorithmic power of AMC that is originally tailored for $\hat{Y}$-problems, but now recalibrated for $\hat{\tau}$-problems. By predicting (imputing) potential outcomes, the scholar benefits from flexible models, but as a side effect of this statistical practice—of combining AMC for causal inference—the original distinction between $\hat{\beta}$ and $\hat{Y}$ has dissipated.

Predicting potential outcomes constitutes one necessary step in calculating counterfactuals, and thus, individual-level causal effects (Pearl, 2009). The definition of potential outcomes and counterfactuals are closely related, but they refer to different scenarios. Potential outcomes refer to a scenario where the treatment assignment has not been made yet, and thus, before the treatment has been assigned, an individual has two potential outcomes, $Y_i^1$ and $Y_i^0$. Counterfactuals refer to a scenario where the treatment has been assigned, but the scholar imagines what the outcome would have been had the treatment assignment been different. A counterfactual exists after the treatment has been assigned. For example, if the individual was assigned the treatment, and therefore his or her factual outcome equals the potential outcome under treatment $Y = Y_i^1$, then this individual’s counterfactual outcome is $Y_i^0$. Thus, counterfactuals enable retrospective reasoning of the form, ‘What if I had acted differently, would the result turned out the same’ (Pearl, 2019).

To calculate the value of a counterfactual, scholars must make assumptions about the noise (error) variables, $U$, in a causal system (DAG) (Pearl, 2009). These variables represent any exogenous events—occurrences that are only indirectly relevant to the causes and effects of a DAG—that induce variations across individuals, and thus when this noise is known, it uniquely determines everyone’s values in the data. These variations represent
all factors that are particular to each individual, yet they are not necessary to the DAG, and thus, they are not always explicitly specified. For example, although an individual’s genetics calibrate physiology and thus nutritional intake, in famine situations, genetics do not directly add to explaining famine outcomes. Although in large samples, these variations cancel each other out when calculating the average treatment effect, $\tau = E[Y_i^1 - Y_i^0]$, these variations are key to calculate individual-level treatment effect, $\tau_i = Y_i^1 - Y_i^0$.

As the context $X$ and variations $U$ jointly determine the exact conditions when the individual took a treatment versus did not take it, we need to know the structural causal model of the causal system and the distributions of both $X$ and $U$ to calculate $\tau_i$. A scholar can only gather these quantities when the DAG and its structural causal relationships are known (Pearl, 2009). When they are known, counterfactuals enable probability expressions such as $P(Y^{W=1} | W = 0, Y^{W=0})$, standing for ‘the probability of observing the potential outcome $Y^{W=1}$ had the exposure $W$ taken the value 1, given that we actually observed the outcome with exposure $W = 0$ (that is $Y^{W=0}$).’ For example, in our famine case, this probability can refer to a specific Bengali farmer: would the farmer have survived $Y^{W=1} = 1$ had the Bengali government distributed food coupons (entitlements to food) to farmers $W = 1$, given that this farmer actually starved to death $Y^{W=0} = 0$ and did not receive coupons $W = 0$ (Daoud, 2017). Although counterfactuals necessarily rely on stronger assumptions than calculating average effects, they present an exciting path for applied domains, such as personalized medicine (Gottesman et al., 2019), precision agriculture (Bauer et al., 2019), and public policy (Balgi, Peña, & Daoud, 2022a, 2022b). Pearl takes this statement one step further by arguing that “Advances in graphical and structural models have made counterfactuals computationally manageable and thus rendered causal reasoning a viable component in support of strong AI” (Pearl, 2019, p. 1).

Predicting counterfactuals and potential outcomes enables an inductive analysis of treatment heterogeneity (Athey et al., 2019; Imai & Ratkovic, 2013; Künzel et al., 2018). While average treatment effect analysis focuses on the aggregated effect of an exposure on a population, effect heterogeneity analysis focuses on more granular effects, the group-specific effects disaggregated by subpopulations (Shiba et al., 2021; Shiba, Daoud, Hikichi, et al., 2022; Shiba, Daoud, Kino, et al., 2022). For example, although a famine or an economic crisis is likely to affect an entire country adversely, some combination of socioeconomic factors may protect certain groups better than others (Daoud & Johansson, 2020). In DMC, scholars tend to capture such effect heterogeneity with the help of interaction models. Because these models are parametric, they often need a scholar to specify the product terms explicitly. If scholars hypothesized that ethnicity moderated the effect of entitlements in explaining famines, then they would deductively specify the following model,

$$FAMINES_j = c_0 + \beta_1 ENTITLEMENTS_i + \beta_2 ETHNICITY_i + \beta_3 ENTITLEMENTS_i \cdot ETHNICITY_i.$$  

If the parameter $\beta_3$ is statistically significant, then that is evidence for treatment heterogeneity.

However, as shown in our T-learner example, an ML model for treatment heterogeneity does not require such explicit specification. In an ML model for causal inference, the effect heterogeneity for our famine example is defined as the conditional average treatment effect (CATE), $\tau(x_i) = E[FAMINES_i | X = x_i]$ where
While a parametric model tests a specific parametric interaction, this ML model searches over the joint conditional distribution of \( p(FAMINES_i \mid X = x_i) \) for group-specific causal effects, where these groups are defined by \( X = x_i \). Again, because of the flexibility of ML models, not only are HMC practices capturing the average treatment effect more robustly but these models find effect heterogeneity automatically.

### 3.1.2. ML predicts propensity scores and similar metrics

In the second combination, scholars apply ML in the service of commonly used causal methods. Even if the statistical model of interest is a parametric model where \( \hat{\beta} \) imprints the causal effect of interest, ML can service this model in an initial estimation step. While many parametric approaches that rely on two or more estimation steps can benefit from such a service, instrumental variables methods (Belloni et al., 2014, 2018; Carrasco, 2012) and propensity score models (Alaa et al., 2017; Ju et al., 2017; Wyss et al., 2017) have benefited most. An instrumental variable method is an identification and estimation technique used in a situation when the targeted causal relationship between the treatment \( W = w \) and an outcome, \( Y = y \), is confounded by an unobserved event, \( C = c \). In this situation, an instrument \( Z = z \) disentangles the variation of \( C \) on \( Y \) from the variation of \( W \) on \( Y \), thereby capturing the targeted causal effect. A key assumption is that the instrument affects only the treatment, and that this affect mimics a random coin flip (Morgan & Winship, 2014). Regressing the treatment on the instrument, a statistical model targets the variation of \( w \) that is determined by the instrument and not the contaminated contaminating effect of the confounder \( c \). The instrumental variable method comes in different versions but the basic one is the two-stage instrumental variable procedure. This procedure consists of the following stages: the first stage estimates \( w_i = c_0 + \gamma z_i + e_i \), where \( z_i \) is the instrument and the second stage \( y_i = c_1 + \beta \hat{w}_i + e_i \) uses the predicted (targeted) version of \( \hat{w} \) instead of the original \( w \) (contaminated). As \( z \) mimics a coin flip, \( \hat{w} \) transforms the treatment assignment to an as-if random event. As the first stage is a prediction problem, recent literature developed a variety of different extensions from deep-instrumental variable approaches to sample-splitting procedures strengthening the capabilities of this framework (Belloni et al., 2018; Hartford et al., 2016).

Similarly, instead of relying on logistic regression to estimate a propensity score model and running the risk of overfitting, new methods utilize AMC-type procedures and algorithms to estimate propensity scores (Lee et al., 2010). Then, these scores are used for downstream causal estimation methods, such as inverse probability weighting. Many current methods combine the best of two worlds by predicting the treatment propensity and evaluating an outcome (regression) model (Chernozhukov et al., 2018; van der Laan & Rubin, 2006; Nie & Wager, 2018; Schuler & Rose, 2017; Sverdrup et al., 2020).

### 3.1.3. ML facilitates interventional and sequential decision making

In the third method combination, HMC scholars use ML for policy optimization for dynamic sequential decision-making (Russo et al., 2018). By dynamic, we mean a situation where a treatment assigned at time \( t \)
has a causal effect not only on the outcome at time $t + 1$, but also on subsequent treatment decisions (Hernan & Robins, 2020). For example, a doctor treating a cancer patient seeks to identify the optimal treatment with the least amount of pain (Gottesman et al., 2019; Murphy, 2003). What and how much dosage this doctor decides to inject into the bloodstream of her patient at time $t$ will not only affect the patient’s pain level in the next sequence but also affect the doctor’s set of options in future sequences. Consequently, this sort of sequential decision-making problem translates to finding the optimal policy with the desired causal effect with as few treatment steps as possible (Murphy, 2003).

On the one hand, the problem of ‘searching-over-potential treatments (or actions, policies) to find the optimal effect’ aligns with DMC’s ambitions of identifying a causal effect. On the other hand, this problem also has a predictive structure similar to those AMC problems of algorithms playing computer or board games (Russo et al., 2018). Consider DeepMind’s reinforcement-learning algorithms AlphaStar and AlphaGo that are able to select the optimal decision at time $t$ that predicts the best chance of eventually winning the game. These algorithms are purely predictive, lacking any causal component. Yet they work. Based on the predictive power of reinforcement-learning algorithms, HMC-influenced scholars combine these algorithms and causal inference (Zhang & Bareinboim, 2020). One way of achieving these combinations is by estimating the causal effect of a particular decision and predicting the final outcome for a sequence of similar decisions. The algorithms achieve this complex task by using similar principles from the first way in which ML supports causal inference—imputing potential outcomes. Reinforcement algorithms impute potential outcomes for many possible sequences, and then based on these synthetic data, they select optimal decisions. Scholars have also combined these algorithms with wearables and sensing technologies, thereby embedding sequential medical interventions directly into patients’ daily lives (Liao et al., 2020). Other scholars explore how reinforcement algorithms can be used to find optimal economic policies (Kasy, 2018; Zheng et al., 2020).

### 3.1.4. ML discovers causal systems

In the fourth method combination, scholars apply ML for causal discovery. Many social-scientific domains lack robust theories about how events in that domain are causally connected (Swedberg, 2017). The lack of such theories leads to imprecise DAG representations of the causal system of interest. To remedy this lack and to fuel causal theorizing, causal-discovery algorithms suggest DAGs for further analysis (Jaber et al., 2018; Peters et al., 2017). Assuming the existence of observational data that represents all key variables of a causal system, these algorithms search over the covariate space to find and suggest DAGs in a data-driven way (Spirtes et al., 2001).

The causal-inference toolbox offers different algorithms for suggesting DAGs (Glymour et al., 2019). An independence-based algorithm tests for permutations of conditional and unconditional independence among variables $X_1, X_2, \ldots, X_k$, using their joint distribution $p(X_1, X_2, \ldots, X_k)$. In a data set containing only two variables, if that algorithm assesses that these variables are likely independent $X_1 \perp X_2$, then it assigns a low probability that these two variables are causally connected. Because they are independent, their
joint distribution factorizes into \( p(X_1, X_2) = p(X_1) p(X_2) \), and the algorithm abstains from drawing an edge between these two variables. For three variables, if the algorithm finds that \( X_1 \) and \( X_2 \) are independent conditional on \( X_3 \), that is, \( X_1 \perp X_2 | X_3 \), then this conditional independence yields three mutually exclusive interpretations. First, this finding could suggest that \( X_3 \) is a mediator because it lies on the path between \( X_1 \) and \( X_2 \), like the following. The joint distribution of this path factorizes into \( p(X_1, X_2, X_3) = p(X_1) p(X_3 | X_1) p(X_2 | X_3) \). Second, \( X_3 \) could be a mediator in the other direction \( X_1 \leftarrow X_3 \leftarrow X_2 \) with a joint distribution of \( p(X_1, X_2, X_3) = p(X_2) p(X_3 | X_2) p(X_1 | X_3) \). Third, that conditional independence suggests also that \( X_3 \) could be a common cause, \( X_1 \leftarrow X_3 \rightarrow X_2 \) with a joint distribution that factorizes into \( p(X_1, X_2, X_3) = p(X_3) p(X_1 | X_3) p(X_2 | X_3) \). To discern among these three interpretations, the algorithm requires more information. If the data set has additional variables, the algorithm continues adjusting the edges and filtering the most likely DAGs from the last plausible based on other independencies.

No matter the data matrix’s width or length, causal discovery suffers from at least three limitations. First, because many discovery methods rely on conditional-independence tests, perturbations in the data—either through sampling or measurement error—may flip the causality direction, especially if the true causal effect is small (Shah & Peters, 2020). While conditional-independence tests are relevant for many statistical practices (Dawid, 1979), these tests form the foundation for many discovery algorithms, and the success of these algorithms depends on the capacity of such tests (Shah & Peters, 2020). Second, causal discovery assumes that all relevant variables about the causal system of interest are measured (Robins & Wasserman, 1999). That is a strong assumption, required because competing social theories \( T_k \) will likely stipulate different data representations and measurements. These requirements will likely not be fulfilled for even a modestly complex causal system.

Third, even if all variables were measured and conditional-independence tests were unbiased, causal-discovery algorithms rarely find one optimal DAG, but many candidate DAGs that are equally well suited to represent the same causal system given the data (Peters et al., 2017). The observed data can only do so much. Nonetheless, combined with domain knowledge or randomized-control trials, scholars continue the filtering of plausible and implausible DAGs. This combination of machine-suggested DAGs and human-domain knowledge of the causal system have proven fruitful in, for example, genetics. Because of the complexity of how genes interact and regulate each other, scholars are yet to precisely determine the causal direction of any genetic system. Causal-discovery algorithms have proven useful in suggesting a causal representation of how genes regulate each other (Glymour et al., 2019). While causal discovery is a vibrant field of research with promising contributions, applied-social science is yet to evaluate its usefulness.

In sum, causal inference is benefiting from ML in at least four different ways. As discussed, by a mixture of DMC and AMC practices, a new set of synthesized statistical practices have emerged. This new culture, HMC, weaves prediction and inference into synthesized procedures. ‘ML for imputing potential outcomes’ is perhaps
the clearest confirmation of the existence of HMC. Here, the goal is to capture $\hat{\tau}$, but through ML imputations the traditional distinction of $\hat{Y}$ and $\hat{\beta}$ has become superfluous.

3.2. ML for Data Acquisition

Ideally, scholars would be able to measure all the necessary variables that represent the causal system of interest. The primary and minimum variables for causal inference in observational studies are a treatment $W$, an outcome $Y$, and a confounder $C$. Figure 3 shows a DAG containing the basic set of variables for causal analysis in observational settings. If either the treatment or outcome is unobserved, quantifying their causal connection $\tau$ is impossible; if only the confounder is unobserved, statistical models will produce biased results for $\tau$. Because a confounder is a variable that affects both the treatment and the outcome, the magnitude of this bias depends on how strongly this confounder affects both of these two variables (Hernan & Robins, 2020). For any scientific domain, therefore, measuring high-quality data about the causal system of interest remains a crucial task.

![Figure 3. The basic set of variables for causal analysis in observational settings.](image)

The second practice of HMC mobilizes AMC-type algorithms to support this task. Table 2 defines the key characteristics of these measurement practices. While in DMC-type practices, scholars rely more on analog methods to measure data and less on digital ones, in AMC, this emphasis is reversed (Salganik, 2017). An analog-measurement method relies predominantly on humans to measure and structure information about causal events. Conducting surveys, interviews, performing experiments, and conducting ethnographic studies are examples of analog methods. A digital-measurement method is any method that also relies on algorithms to
extract data from structured or unstructured digitized sources. A *digital* source is a piece of information existing as ‘1s’ and ‘0s’ on a computer. By *structured*, we mean information that exists as a tidy data matrix in well-defined variables and values. Conversely, *unstructured* information has yet to be preprocessed for a meaningful structure. Digitized historical archives is one example of unstructured information (Salganik, 2017). Processing a large amount of structured and unstructured information is one of the tasks where ML excels (Blei et al., 2003).

Mobilizing research assistants to code up the political content of archival policy documents exemplifies an analog method; training natural language processing (NLP) algorithms to do the same thing is a digital method (Grimmer & Stewart, 2013). Likewise, in famine research, employing assistants to analyze geographical maps to code events of drought is an analog method; training image-recognition algorithms to detect drought in satellite images is a digital method (Daoud et al., 2021; Mahecha et al., 2020). Each method has its strengths and weaknesses. Because analog methods rely on humans, these methods are better tuned to measure sensitive content, but they are slower and more costly. While digital methods still require human supervision for training data and interpreting the content of unsupervised results, they are still faster and cheaper when applied to large data sets (Salganik, 2017).

In the digital era, several additional petabytes of data are made available every year. Equipped with DMC logic for sampling, HMC scholars use AMC practices to efficiently sift through these data for measurement. Although the role of digital methods is considerable in HMC, analog methods remain an essential part of HMC practices. While unsupervised algorithms reduce high-dimensional data (e.g., an archival document) to low-dimensional representations (e.g., a topic-model distribution), scholars still have to interpret these low-dimensional representations manually—what Chang et al. (2009) metaphorically compared to reading tea leaves. Unsupervised ML consists of algorithms that reduce dimensionality in a set of covariates $\mathbf{X}$, without any reference to a specific outcome, $\mathbf{Y}$, thereby requiring more human supervision for interpretability. Because supervised ML consists of algorithms tailored to predict a prespecified outcome, $\mathbf{Y}$, using $\mathbf{X}$ as input, these algorithms require high-quality labeled data prepared by scholars. Often this labeling of data relies on surveys or qualitative coding of digital sources, based on the scholars’ expertise. The gain of supervised ML is that the algorithm is more effective in generalizing to unlabeled data (Hastie et al., 2009).

When applying ML to process digital sources, scholars are conducting a form of data measurement. Instead of asking people directly about their material living standards, scholars let a machine capture these standards from digital sources (Daoud et al., 2021; Jean et al., 2016). A limitation of using ML to conduct such measurement is the added error in the data processing.

The total-survey-error approach helps characterize the sources of error in traditional surveying (Groves & Lyberg, 2010). Analog surveys pose specific challenges arising from formulating questionnaires, conducting interviews, and other disturbances affecting measurements. This approach provides a framework to characterize what these sources of error are when a scholar measures an event $\mathbf{X}^*$ and quantifies it as a
variable in the data $X'$. Although different errors exist, they categorize either as bias (systematic error) or variance (random error). **Systematic error**, $e_s$, is any information that shifts the sample estimate $X'$ away from the true value $X^*$ in a consistent manner. Inaccurately calibrated instruments usually cause such shifts. An instrument is a means for acquiring information about the event of interest. For example, poorly phrased wordings in a questionnaire (the instrument) will lead to over- or underreporting of a respondent’s behavior. Systematic error is reduced by improved calibration of the instrument. **Random error**, $e_r$, is the natural variation arising from sampling procedures that affect the accuracy of the instrument. Random errors cancel each other out as the sample size increases because negative deviations for one individual $X'_i$ eventually cancel out by a positive deviation for another $X'_j$. Putting these two sources of error together, a survey will always be an imperfect representation of any variable in a causal system, as described by the following formula, $X^* = X' + e_s + e_r$.

While scholars can improve their instruments and surveying execution to reduce these two sources of error, they will eventually hit a limit where they start trading bias for variance or vice versa. Both analog and digital methods suffer from the same limitations, yet digital methods inject additional errors (Salganik, 2017).

Because many supervised ML algorithms rely on the analog data source for training samples, these algorithms can only recreate imperfect representations of $Y'$ and not of the true event $Y^*$. This imperfection arises from training and testing the algorithms (Hastie et al., 2009). Famine and poverty research serves us as an illustration. Many scholars use analog methods to measure people’s living conditions, $Y^*$, by surveying people’s income or material assets, $Y'$. In designing and executing their surveys, scholars encounter both systematic and random errors, $e_1 = e_{1s} + e_{1r}$, that will distort their sample estimate, $Y'$. These errors are additive if there is no interaction between systematic and random errors. Nonetheless, $Y'$ is the best analog representation a scholar can produce to capture $Y^*$.

Subsequently, suppose that another group of scholars is aiming to speed up the surveying of poverty by combining analog and digital sources such as satellite images (Blumenstock et al., 2015; Daoud et al., 2021; Jean et al., 2016; Yeh et al., 2020). These images reveal the living conditions of people, as they appear from the sky. So, this group collects $Y'$ as a training sample $y'_1, y'_2, \ldots, y'_n$ (analog data) for their ML algorithm, and satellite archives (digital data), $S$. Their goal is to train an algorithm $f$ to measure (predict) income $\hat{Y}'$ from the pixel features of these satellite images $S$. This procedure constitutes defining $Y' = f(S) + e_2$, where $f(S) = \hat{Y}'$ and an error composed of $e_2 = e_{2s} + e_{2r}$, assuming again additive errors. Although this ML-satellite approach to measure poverty is faster than letting humans survey poverty, each new ML step added to represent $\hat{Y}'$ continues eroding at the human survey produced $Y'$. This $Y'$ is already an imperfection of $Y^*$, and each ML step induced new systematic and random errors. This erosion is expressed in the following mathematical relationships for one analog and one digital measurement,
Systematic and random errors compound from the analog (i.e., $e_{1s} + e_{1r}$) and digital methods (i.e., $e_{2s} + e_{2r}$). The more algorithmic transformations added on top of the first procedure $\hat{Y}'$, the more these errors are likely to propagate, continuing to corrode $Y^*$, and thereby aggravating the measurement error.

Although measuring events digitally induces an additional error, digital approaches are usually faster and less costly. These two advantages have prompted HMC-influenced scholars to new sources of data to impute the missing data to populate their representation of the causal system of interest—a procedure that Bareinboim and Pearl (2016) formalized under the name *data-fusion problems*. For example, scholars use topic models or other representations to summarize text to capture confounding (Åkerström et al., 2019; Blei et al., 2003; Blei, 2012; Daoud, Jerzak, & Johansson, 2022; Egami et al., 2018; Mozer et al., 2020; Roberts et al., 2016); they process images to measure outcomes (Daoud et al., 2021; Jean et al., 2016) or confounding (Jerzak et al., 2022, in press); they assemble corpora of health records to record patients’ background (Hsu et al., 2020); they use video and audio to capture other representations to facilitate scientific inquiry (Tarr et al., 2022; Knox & Lucas, 2019). These new measures are then used for downstream causal inference tasks.

However, a critical challenge of using digital sources is when the same source is used two or more times for measuring multiple variables, or when information about one variable leaks into another variable. Daoud, Jerzak, & Johansson (2022) calls this *treatment leakage* when it involves measuring the confounding variable but treatment information leaks into the measurement of this confounding variable. This leakage leads to post-treatment bias. Thus, refining analog and digital measurement approaches will remain a crucial part of the scientific endeavor.

### 3.3. ML for Theory Prediction

A vital goal of the scientific endeavor is to explain not only past observations of how an event causes another, $W \rightarrow Y$, but also predict future instances of these causal events (Gelman & Imbens, 2013; Watts, 2014). This goal translates to testing a theory’s predictive power: *theory prediction*, for short (Freedman, 1991; Kleinberg et al., 2017; Marini & Singer, 1988; Peysakhovich & Naecker, 2017; Salganik, Lundberg, et al., 2020; Watts, 2014). One test of a scientific theory, $T_k$, and its corresponding DAG, $G_k$, is to evaluate the amount of statistical support it received in data, focusing on causal effects $\tilde{f}$. Another test of the usefulness of a scientific theory is to evaluate how predictive $T_k$ is for other populations (Billheimer, 2019). While Section 4.1 describes how HMC scholars test, discover, and build theories on causal relationships, this section shows how
these scholars evaluate the predictive performance of their theories. Table 2 shows the key characteristics of this HMC practice.

Although Breiman (2001a) argues in favor of prediction as a tool for fueling the scientific endeavor, he remains unclear about how exactly prediction provides that support. We define a DMC-prediction as a recreation of an event \( Y \) that is as similar as possible to the true event yet not observed but generated by a validated statistical model specified under \( T_k \) and its DAG, \( G_k \). In our famine example, the statistical model \( f_1(\text{FOODSCARCITY}_i) \) representing a Malthusian DAG \( G_1 \) constitute one such validated model. A DMC prediction of famines is then \( \hat{Y}_{G_1} = f_1(\text{FOODSCARCITY}_i) \), for famine events \( Y_i \) with characteristics \( \text{FOODSCARCITY}_i \), for cases \( i \) that the statistical model \( f_1 \) has not been observed before. In contrast, an AMC prediction is a recreation of an event that is as similar as possible to the true event yet not observed, but it is not necessarily conditioned on any validated causal model. In its distilled form, an AMC prediction is a pure prediction problem that uses an algorithm and data source (Kleinberg et al., 2015). For example, an AMC prediction of our famine example involves collecting any input variables that carry some association to the outcome (famines), training an ML model on these data, and then evaluating this model’s predictive power on a held-out set (Okori & Obua, 2011). AMC predictions constitute a horserace among candidate algorithms and data competing for the best predictive performance—Kaggle-style competitions. DMC predictions are also horseraces but only among scientific theories and their statistical representations (Daoud et al., 2019).

HMC predictions are DMC predictions, but because HMC predictions rely on ML, they submit its prediction practices to the principles dictated by AMC. The two essential principles of AMC-prediction practices are the use of regularization and evaluation using held-out samples. These two principles minimize overfitting that will likely come about when scholars attempt to squeeze more variables into their models or tweak their models’ functional form to fit the data better in-sample. Additionally, as future observations are unobserved in the present, held-out samples stand in for these missing observations (Risi et al., 2019). Highly predictive theories have a small difference, \( \epsilon_{\hat{Y}_{G_k}} \), between what the theory says will happen \( \hat{Y}_{G_k} \) and what eventually did happen, \( Y \), such that \( \epsilon_{\hat{Y}_{G_k}} \approx Y - \hat{Y}_{G_k} \). Two theories compete for superiority in explaining \( W \to Y \), generate their respective predictions on new data, and then evaluate their predictive performance. They do so by using their respective DAGs, \( G_k \), which defines how context \( X \) affects \( W \to Y \). The theory scoring the smallest \( \epsilon_{G_k} \), wins a theory-prediction contest.

A second way to test a theory’s predictive power is to evaluate whether the identified causal effect \( \hat{\tau}_{G_k} \), based on a DAG \( G_k \), found in one population can be verified in another, preferably a randomized control trial when feasible. In computer science, these verifications are called transportability of results (Pearl & Bareinboim, 2014), and they are also known as transfer learning, life-long learning, and domain adaptation (Chen & Liu, 2018; Johansson et al., 2019); in statistics, and elsewhere, they go under the name of generalizability or external validity (Deaton & Cartwright, 2016). Some scholars call these verifications forward-causal questions when the event is generalized to lie in the future (Gelman & Imbens, 2013). For example, finding that a social
policy is working as intended for a population (e.g., the poor) in one county, scholars may ask how well this policy generalizes for the same county but for future populations. The more populations for which \( \hat{\tau} \) exists and has a similar value, the more support a theory receives.

To systematically evaluate the predictive performance of different theories’ predictions (i.e., \( \hat{Y}_{G_k} \)) and causal claims (i.e., \( \hat{\tau}_{G_k} \)), scholars require a common task framework (Donoho, 2017). This framework is a set of principles defining how predictive performance is scored and evaluated. Its purpose is to ensure the comparability of results. Such a framework follows at least three principles. First, scholars require a publicly available training data set for which they may operationalize their respective scientific theories \( T_k \), articulate their causal assumptions \( G_k \), and formulate their statistical models \( f_{G_k} \). This data set must be sufficiently rich to accommodate many different plausible theories \( T_k \). Second, as in any competition, scholars have to make themselves known to each other and agree to the rules of the competition. These rules must at least specify one estimand (quantity of interest) and how scholars’ estimators will be evaluated (e.g., minimizing mean squared error). Third, the competing scholars have to designate a scoring referee for which they can submit their estimators. This referee automatically and objectively checks each scholar’s estimator against a held-out data set that has been kept secure behind a firewall during the competition. The referee presents the results and estimators transparently, enabling the competing scholars to reproduce each other’s results, thereby enhancing scientific learning.

Much of the success of AMC is supported by well-functioning common task frameworks (Donoho, 2017). Often there are clear outcomes \( Y \) defined, and transparent procedures for scoring each competitor prediction, \( \hat{Y} \). For example, much of the development of image-recognition algorithms owns its success to the publicly available data source, ImageNet (Deng et al., 2009). It gave deep-learning scholars a common benchmark, which new algorithms could be tested against. In DMC, setting up a similar infrastructure is challenging because the estimand is often a causal effect, \( \tau \), which is per definition, unobserved in the data. As neither the scoring referee nor the scholars know the true \( \tau \), there is no way to score contending estimators. This implies that a common task framework for causal problems has to be complemented with additional assumptions for the competition to work. One way of handling this insufficiency is to combine observational data with randomized control-trial data (Lin et al., 2019). As RCTs require the least assumptions, one target estimand is the average treatment effect in RCT, \( \tau_{RCT} \). However, this target will only work if the RCT has high internal and external validity. Another way of handling this insufficiency is to rely on simulations in which the causal effect is known exactly, \( \tau_{sim} \). Although that solves the problem of the target estimand, it introduces the problem that the data is merely an artificial representation of a causal system.

While establishing a common task framework to evaluate causality remains a challenge in many disciplines—especially in the social sciences—several common task frameworks focus on an observable estimand: predicting outcomes \( Y \). For example, the Fragile Families Challenge is a scholarly mass collaboration tailored to predict six life outcomes for children age 15 (Salganik et al., 2019; Salganik, Lundberg, et al., 2020;
Salganik, Maffeo, et al., 2020). These outcomes are child grade point average (GPA), child grit, household eviction, household material hardship, caregiver layoff, and caregiver participation in job training. This challenge received 437 scholarly competitors (of which some worked in teams) that resulted in 160 valid submissions. All these submissions were evaluated using means squared error in held-out data. In the social sciences, this challenge is among the first to devise a common task framework for the advancement of science (Meng, 2020).

4. Conclusion

To recap, social scientists regard predictive statements and causal inferences as two distinct research problems (Watts, 2014). To a large extent, this distinction follows the fault line of the two cultures of statistical modeling (Breiman, 2001a). The data modeling culture (DMC) is the modus operandi in applied social science; the algorithmic modeling culture (AMC) dominates computer science, engineering, and many policy practices. DMC research tends to favor causal inquiry but is limited by being caught in model validation; AMC embraces predictive inference but is not well tailored for the scientific method. Nonetheless, while pure causal and predictive inferences have their place in the social sciences, our article has shown that there is a third kind of problem that synthesizes predictive and causal practices to the extent that it is hard to tell them apart (Daoud & Dubhashi, 2021). Given the new scientific opportunities and challenges arising in the digital era, methodologists have developed various techniques fueling the scientific endeavor (Athey & Imbens, 2019; Imai, 2018; Kino et al., 2021; Lazer et al., 2020; Pearl, 2019). Through these developments, a new modeling culture has evolved that has mutated from DMC and AMC: the hybrid modeling culture (HMC).

This article has identified the main characteristics of HMC and showed how it synthesizes components of DMC and AMC under the umbrella of explanatory (causal) social-scientific research (Boudon, 2005; Hedström & Ylikoski, 2010; Lundberg et al., 2021; Marini & Singer, 1988; Merton, 1968; Risi et al., 2019; Watts, 2014). First, the overarching aim of HMC is to further the production of knowledge. HMC copies this aim from DMC, where the overarching goal is to explain how $X$ causally affects $Y$. Scholars achieve this goal by assuming a causal system, $G_{T_k}$ stipulated by a substantive theory $T_k$. Under these assumptions, they test competing explanations against data, refute, revise, or update theories depending on these tests’ results. Second, HMC does not restrict itself to the commonly used statistical models offered by DMC for statistical and causal inference but incorporates the range of powerful algorithms offered by AMC (Bail, 2017; Lazer et al., 2020; Nelson, 2020; Turco & Zuckerman, 2017; Watts, 2017). As scholars combine AMC-type algorithms for DMC-inspired inference (Lundberg et al., 2021; Molina & Garip, 2019), they erode the traditional distinction between prediction $\hat{Y}$ and statistical inference $\hat{\beta}$. This erosion recasts the scientific problem into $\hat{\mathcal{P}}$ problems—identifying ways to impute (predict) potential outcomes $Y^1$ and $Y^0$.

A possible criticism of HMC is that because it has the same goal as DMC, it would suggest that it is not all that separate of a culture from DMC. While these two cultures have the same goal, they use different practices to achieve that goal. Another way to appreciate the differences among the statistical cultures—HMC, DMC, and
also AMC—is to recast the issue of statistical *culture* as an issue of methodological paradigms (Kuhn, 2012). A *methodological paradigm* is a way of looking at the world of data and models. Within the paradigm of DMC, scholars would be violating fundamental principles of statistical modeling if they relied on ML for causal inference. Often, the (causal) estimands of DMC is a parameter $\beta$ in a prespecified model; the causal estimand of HMC is $\tau$. Some of the debates that raged between scientific camps within the field of causal inference—for example, those between the perspectives of Judea Pearl and Donald Rubin—are likely to have their roots in the paradigmatic differences between DMC and HMC (Imbens, 2020). Some of those debates pertain to the use of DAGs, but several of them are about what estimands are scientifically valuable to target in the first place. This targeting implies also that DMC is more stringent about using traditional causal estimation methods, while HMC is lax. Even if HMC is not associated with a particular method, we have shown that it tends to use AMC-type methods. Not because these ML algorithms are fancy, but because they are suitable for the age of data science and for scientific inquiry.

It is a vindication for HMC—but perhaps also a historical irony—that Breiman proposed the use of the random forest as a hallmark of AMC (Breiman, 2001a, 2001b), but then two decades later, that same algorithm has been adapted for HMC-style causal inference. At the time of writing, there are several derivatives of the random forest such as the generalized random forest (Athey et al., 2019), Bayesian Additive Regression Trees (Hahn et al., 2020; Hill, 2011), and other tree-based methods for effect heterogeneity (Brand et al., 2021).

Hybrid modeling culture informs a better social science than AMC or DMC. Embracing the logic of HMC will likely enable scholars to venture beyond current frontiers with deeper confidence. First, loosening the dependence on DMC-type linear models will likely produce more robust research. Models will be less dependent on researchers’ discretion in specifying, tweaking, and tinkering with statistical models. Machine learning algorithms are not entirely immune to tinkering, but the processes of cross-validation and regularization provide some safeguarding against cherry-picking results. Second, and relatedly, the use of ML for causal inference may help alleviate the replication crisis in the social sciences. Successfully replicating a study has several dimensions, but what HMC can assist with is precisely the use of more robust models via not only cross-validation but also regularizing ML algorithms. As previously mentioned, regularization forces an algorithm to generalize better to other data than DMC-type models. Third, embracing the logic of HMC implies that scholars will be better equipped to navigate the use of analog and digital sources. When all the sufficient data (e.g., a survey) is available in analog form, then rely on these data only; but when such analog data are insufficient for the causal estimand of interest, an HMC-informed scholar will be able to better navigate statistical issues than a DMC scholar. The HMC-informed scholar will with deeper competence combine analog data with digital sources to use ML to measure what is missing.

Fourth, HMC will likely inspire social scientists to increase the precision of their theories. Currently, many theories abstain from defining all the structural relationships in the causal system because of the inherent causal complexity of the social world. Nonetheless, combining human-assisted causal discovery algorithms will
enable researchers to process large data, and to start engaging in translating their theories into well-defined structural relationships. Defining structural relationships means characterizing not only the relationship of interest, that is, between the cause and effect, \( W \rightarrow Y \), but also characterizing how the context variables, \( X_1, X_2, \ldots, X_z \), are related among each other and to \( W \rightarrow Y \). For example, a theory defining exactly how the social environment interacts with human genes for explaining students’ university grades is a complex task because scientists do not have strong theories about all the causal structural relationships between genes and social environment (Beauchamp, 2016; Courtiol et al., 2016). Using causal discovery algorithms is a viable way of starting to create some order in these complex data and thereby refining social theories.

Hybrid modeling culture gives rise to new methodological challenges and thus inspires at least three directions for future methodological research. Standard errors are the foundation of DMC-type of inference, but they measure only sampling variability. Quantifying sampling variability is insufficient for HMC, as there are additional uncertainties threatening HMC-type inference. Because HMC scholars use multiple data sources and models, and in different phases, they need to capture additional uncertainties reflecting this multitude. They need to handle a trinity of inference: multisource, multiphase, and multiresolution (Blocker & Meng, 2013; Li & Meng, 2021; Meng, 2014).

**Multisource** refers to the increasingly common situations where a single study uses data sets from different sources but with varying quality (Meng, 2018). Because of this variation, *big* does not imply better, and a large amount of biased data can do far more damage than smaller data sets because they can lead us to be overly confident in erroneous results. In the case of using satellite images for measuring poverty, data comprises household surveys, satellite images, nightlight data, and other sources (e.g., ImageNet for transfer learning). These data are collected from different continents and years and thus are plagued by sampling variation and biases comprising different satellite technologies, seasonality, and changing survey definitions (Burke et al., 2021). A critical question is then how scholars may account for these multisource uncertainties beyond merely producing standard errors in DMC-type statistical inference.

**Multiphase** refers to the common practices where data are typically collected, preprocessed, and analyzed sequentially by parties with different goals and access to information and limited communications among them (Blocker & Meng, 2013). As a result, scholars may encounter the multiphase inference paradox: every party engages in a statistically valid process, yet the ultimate output from the collective processes can be statistically invalid due to the uncongeniality among the processes. For example, household surveys are sampled with the statistical aim of representativeness of a country and satellite images for monitoring the planet, yet surveys and images are combined for training remote survey methods to measure health and living conditions (Daoud et al., 2021). But survey data typically suffer from nonresponse, which is often imputed by the survey collectors, and satellite images need to be preprocessed before analysis. Another source of uncertainty is therefore that of data preprocessing, and the question of how scholars may account for these different phases in their analysis.
*Multiresolution* is about the unit of analysis (aggregation) and the resolution of these units, such as measurement frequency and granularities of the features (Li & Meng, 2021). While big data encourage finer resolution analyses (e.g., individualized treatments), there is typically a trade-off between data availabilities and resolution levels: the higher the target resolution the fewer relevant data points. Social-scientific inquiry is about societies, economies, and ecologies, and therefore, various analyses need to be of the appropriate resolution to be relevant for informing and evaluating various local and global policies, yet social science data often do not have the desired resolution. Thus, another set of future questions is about how one can reliably learn from low-resolution data and infer conclusions about the high-resolution target. The mismatch of resolution leads to yet another source of uncertainty.

Even if the trinity of inference can threaten HMC research from several sources of uncertainty, HMC offers a compass for the scientific endeavor in the digital era. The social-scientific endeavor is likely to gain more by moving beyond AMC prediction and DMC inference. HMC supplies a statistical culture that capitalizes on the benefits of ML algorithms while maintaining an eye on the scientific goal: explaining social reality.

**Acknowledgments**

We are grateful for the valuable feedback provided by Xiao-Li Meng, Fredrik Johansson, and offer special thanks to James Bailie for providing that extra feedback round. All remaining errors and omissions are entirely our responsibility.

**Disclosure Statement**

Adel Daoud is funded by the Swedish Research Council.

**References**

Abbott, A. (1988). Transcending general linear reality. *Sociological Theory, 6*(2), 169–186.  
[https://doi.org/10.2307/202114](https://doi.org/10.2307/202114)

Åkerström, J., Daoud, A., & Johansson, R. (2019). Natural language processing in policy evaluation: Extracting policy conditions from IMF loan agreements. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 316–320). Linköping University Electronic Press.

Alaa, A. M., Weisz, M., & van der Schaar, M. (2017). Deep counterfactual networks with propensity-dropout. ArXiv. [https://doi.org/10.48550/arXiv.1706.05966](https://doi.org/10.48550/arXiv.1706.05966)

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering ‘metrics: The path from cause to effect* (with French flaps ed.). Princeton University Press.
Arcaya, M., Raker, E. J., & Waters, M. C. (2020). The social consequences of disasters: Individual and community change. *Annual Review of Sociology, 46*(1), 671–691. [https://doi.org/10.1146/annurev-soc-121919-054827](https://doi.org/10.1146/annurev-soc-121919-054827)

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives, 31*(2), 3–32. [https://doi.org/10.1257/jep.31.2.3](https://doi.org/10.1257/jep.31.2.3)

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*(1), 685–725. [https://doi.org/10.1146/annurev-economics-080217-053433](https://doi.org/10.1146/annurev-economics-080217-053433)

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148–1178. [https://doi.org/10.1214/18-AOS1709](https://doi.org/10.1214/18-AOS1709)

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society, 43*(3), 465–482. [https://doi.org/10.1007/s11186-014-9216-5](https://doi.org/10.1007/s11186-014-9216-5)

Bail, C. A. (2017). Taming big data. *Sociological Methods & Research, 46*(2), 189–217. [https://doi.org/10.1177/0049124115587825](https://doi.org/10.1177/0049124115587825)

Balgi, S., Peña, J. M., & Daoud, A. (2022a). *Counterfactual analysis of the impact of the IMF program on child poverty in the Global-South Region using causal-graphical normalizing flows*. ArXiv. [https://doi.org/10.48550/arXiv.2202.09391](https://doi.org/10.48550/arXiv.2202.09391)

Balgi, S., Peña, J. M., & Daoud, A. (2022b). Personalized public policy analysis in social sciences using causal-graphical normalizing flows. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(11), 11810–11818. [https://doi.org/10.1609/aaai.v36i11.21437](https://doi.org/10.1609/aaai.v36i11.21437)

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences, 113*(27), 7345–7352. [https://doi.org/10.1073/pnas.1510507113](https://doi.org/10.1073/pnas.1510507113)

Bauer, A., Bostrom, A. G., Ball, J., Applegate, C., Cheng, T., Laycock, S., Rojas, S. M., Kirwan, J., & Zhou, J. (2019). Combining computer vision and deep learning to enable ultra-scale aerial phenotyping and precision agriculture: A case study of lettuce production. *Horticulture Research, 6*(1), Article 1. [https://doi.org/10.1038/s41438-019-0151-5](https://doi.org/10.1038/s41438-019-0151-5)

Beauchamp, J. P. (2016). Genetic evidence for natural selection in humans in the contemporary United States. *Proceedings of the National Academy of Sciences, 113*(28), 7774–7779. [https://doi.org/10.1073/pnas.1600398113](https://doi.org/10.1073/pnas.1600398113)

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., & Kato, K. (2018). *High-dimensional econometrics and regularized GMM*. ArXiv. [https://doi.org/10.48550/arXiv.1806.01888](https://doi.org/10.48550/arXiv.1806.01888)
Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies, 81*(2), 608–650. [https://doi.org/10.1093/restud/rdt044](https://doi.org/10.1093/restud/rdt044)

Bhaskar, R. (2008). *A realist theory of science* (Rev. ed). Routledge. [https://lt.ltag.bibl.liu.se/login?URL=http://www.dawsonera.com/depp/reader/protected/external/AbstractView/S9780203090732](https://lt.ltag.bibl.liu.se/login?URL=http://www.dawsonera.com/depp/reader/protected/external/AbstractView/S9780203090732)

Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician, 73*(Suppl. 1), 291–295. [https://doi.org/10.1080/00031305.2018.1518270](https://doi.org/10.1080/00031305.2018.1518270)

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. [https://doi.org/10.1145/2133806.2133826](https://doi.org/10.1145/2133806.2133826)

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*(4), 993–1022.

Blocker, A. W., & Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli, 19*(4), 1176–1211. [https://doi.org/10.3150/13-BEJSP16](https://doi.org/10.3150/13-BEJSP16)

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*(6264), 1073–1076. [https://doi.org/10.1126/science.aac4420](https://doi.org/10.1126/science.aac4420)

Boelaert, J., & Ollion, É. (2018). The great regression. *Revue Française de Sociologie, 59*(3), 475–506.

Boudon, R. (2005). Social mechanisms without black boxes. In P. Hedström & R. Swedberg (Eds.), *Social mechanisms: An analytical approach to social theory* (p. 340 s.). Cambridge University Press.

Brand, J. E., Xu, J., Koch, B., & Geraldo, P. (2021). Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociological Methodology, 51*(2), 189–223. [https://doi.org/10.1177/0081175021993503](https://doi.org/10.1177/0081175021993503)

Breiman, L. (2001a). Statistical Modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. [https://doi.org/10.1214/ss/1009213726](https://doi.org/10.1214/ss/1009213726)

Breiman, L. (2001b). Random forests. *Machine Learning, 45*(1), 5–32. [https://doi.org/10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science, 371*(6535), Article eabe8628. [https://doi.org/10.1126/science.abe8628](https://doi.org/10.1126/science.abe8628)

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Almejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature*.
and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), Article 9. https://doi.org/10.1038/s41562-018-0399-z

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). University of Chicago Press.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2), 383–398. https://doi.org/10.1016/j.jeconom.2012.05.012

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuermans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22 (NIPS 2009)* (pp. 288–296). Curran Associates. https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

Chen, Z., & Liu, B. (2018). *Lifelong machine learning* (2nd ed.). Springer. https://doi.org/10.2200/S00832ED1V01Y201802AIM037

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Costantini, D., & Galavotti, M. C. (1986). Induction and deduction in statistical analysis. *Erkenntnis* (1975-), 24(1), 73–94.

Courtiol, A., Tropf, F. C., & Mills, M. C. (2016). When genes and environment disagree: Making sense of trends in recent human evolution. *Proceedings of the National Academy of Sciences*, 113(28), 7693–7695.

Crupi, V., & Tentori, K. (2010). Irrelevant conjunction: Statement and solution of a new paradox. *Philosophy of Science*, 77(1), 1–13. https://doi.org/10.1086/650205

Danermark, B., Ekstörm, M., Jakobsen, L., & Karlsson, J. (2002). *Explaining society: Critical realism in the social sciences*. Routledge.

Daoud, A. (2017). Synthesizing the Malthusian and Senian approaches on scarcity: A realist account. *Cambridge Journal of Economics*, 42(2), 453–476. https://doi.org/10.1093/cje/bew071

Daoud, A. (2018). Unifying studies of scarcity, abundance, and sufficiency. *Ecological Economics*, 147, 208–217. https://doi.org/10.1016/j.ecolecon.2018.01.019

Daoud, A., & Dubhashi, D. (2021). Melting together prediction and inference. *Observational Studies*, 7(1), 1–7.

Daoud, A., Herlitz, A., & Subramanian, S. V. (2022). IMF fairness: Calibrating the policies of the International Monetary Fund based on distributive justice. *World Development*, 157, Article 105924.
Daoud, A., Jerzak, C., & Johansson, R. (2022). Conceptualizing treatment leakage in text-based causal inference. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 5638–5645). Association of Computational Linguistics. https://aclanthology.org/2022.naacl-main.413

Daoud, A., & Johansson, F. (2020). Estimating treatment heterogeneity of International Monetary Fund programs on child poverty with generalized random forest. SocArXiv. https://doi.org/osf.io/preprints/socarxiv/awfjt

Daoud, A., Jordan, F., Sharma, M., Johansson, F., Dubhashi, D., Paul, S., & Banerjee, S. (2021). Using satellites and artificial intelligence to measure health and material-living standards in India. SocArXiv. https://doi.org/10.31235/osf.io/vf28g

Daoud, A., Kim, R., & Subramanian, S. V. (2019). Predicting women’s height from their socioeconomic status: A machine learning approach. Social Science & Medicine, 238, Article 112486. https://doi.org/10.1016/j.socscimed.2019.112486

Daoud, A., Nosrati, E., Reinsberg, B., Kentikelenis, A. E., Stubbs, T. H., & King, L. P. (2017). Impact of International Monetary Fund programs on child health. Proceedings of the National Academy of Sciences, 114(25), 6492–6497. https://doi.org/10.1073/pnas.1617353114

Darwiche, A. (2017). Human-level intelligence or animal-like abilities? ArXiv. https://doi.org/10.48550/arXiv.1707.04327

Dawid, A. P. (1979). Conditional independence in statistical theory. Journal of the Royal Statistical Society: Series B (Methodological), 41(1), 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

Deaton, A., & Cartwright, N. (2016). Understanding and misunderstanding randomized controlled trials (Working Paper No. 22595). National Bureau of Economic Research. https://doi.org/10.3386/w22595

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Devereux, S. (2007). The new famines: Why famines persist in an era of globalization. Routledge.

DiMaggio, P. J., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics, 41(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004
Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. ArXiv. https://doi.org/10.48550/arXiv.1702.08608

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). *How to make causal inferences using texts*. ArXiv. https://doi.org/10.48550/arXiv.1802.02163

Elder, G. H. (1998). *Children of the Great Depression: 25th Anniversary Edition* (25th anniversary, updated ed.). Westview Press.

Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research* (pp. 245–273). Springer.

Fisher, S. R. A. (1935). *The design of experiments*. Oliver and Boyd.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology*, 21, 291–313. https://doi.org/10.2307/270939

Gelman, A., & Imbens, G. (2013). *Why ask why? Forward causal inference and reverse causal questions* (Working Paper No. 19614). National Bureau of Economic Research. https://doi.org/10.3386/w19614

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10. https://doi.org/10.3389/fgene.2019.00524

Gold, M. K. (2012). *Debates in the digital humanities*. University of Minnesota Press.

Goldthorpe, J. H. (2015). *Sociology as a population science*. Cambridge University Press.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1), 16–18.

Grimmer, J., & Stewart, B. (2013). Text as Data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297.
Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. [https://doi.org/10.1093/poq/nfq065](https://doi.org/10.1093/poq/nfq065)

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with Discussion). *Bayesian Analysis, 15*(3), 965–1056. [https://doi.org/10.1214/19-BA1195](https://doi.org/10.1214/19-BA1195)

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. ArXiv. [https://doi.org/10.48550/arXiv.1612.09596](https://doi.org/10.48550/arXiv.1612.09596)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer New York.

Hedström, P. (2005). *Dissecting the social*. Cambridge University Press. [http://books.google.com/books?id=79IV00LovCYC](http://books.google.com/books?id=79IV00LovCYC)

Hedström, P., & Manzo, G. (2015). Recent trends in agent-based computational research: A brief introduction. *Sociological Methods & Research, 44*(2), 179–185. [https://doi.org/10.1177/0049124115581211](https://doi.org/10.1177/0049124115581211)

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*(1), 49–67. [https://doi.org/10.1146/annurev.soc.012809.102632](https://doi.org/10.1146/annurev.soc.012809.102632)

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. Free Press.

Hernan, M. A., & Robins, J. M. (2020). *Causal inference*. CRC Press.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20*(1), 217–240. [https://doi.org/10.1198/jcgs.2010.08162](https://doi.org/10.1198/jcgs.2010.08162)

Hirshberg, D. A., & Zubizarreta, J. R. (2017). On two approaches to weighting in causal inference. *Epidemiology, 28*(6), 812. [https://doi.org/10.1097/EDE.0000000000000735](https://doi.org/10.1097/EDE.0000000000000735)

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science, 355*(6324), 486–488. [https://doi.org/10.1126/science.aal3856](https://doi.org/10.1126/science.aal3856)

Hsiang, S. (2016). Climate econometrics. *Annual Review of Resource Economics, 8*(1), 43–75. https://doi.org/10.1146/annurev-resource-100815-095343

Hsu, C.-C., Karnwal, S., Mullainathan, S., Obermeyer, Z., & Tan, C. (2020). *Characterizing the value of information in medical notes*. ArXiv. [https://doi.org/10.48550/arXiv.2010.03574](https://doi.org/10.48550/arXiv.2010.03574)
Huber, F. (2022). Confirmation and induction. *Internet Encyclopedia of Philosophy.*
[https://iep.utm.edu/confirmation-and-induction/](https://iep.utm.edu/confirmation-and-induction/)

Tarr, A., Hwang, J., & Imai, K. (2022). Automated coding of political campaign advertisement videos: An empirical validation study. *Political Analysis,* (First View), 1–21. [https://doi.org/10.1017/pan.2022.26](https://doi.org/10.1017/pan.2022.26)

Imai, K. (2018). *Quantitative social science: An introduction* (Illustrated ed.). Princeton University Press.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics,* 7(1), 443–470. [https://doi.org/10.1214/12-AOAS593](https://doi.org/10.1214/12-AOAS593)

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature,* 58(4), 1129–1179. [https://doi.org/10.1257/jel.20191597](https://doi.org/10.1257/jel.20191597)

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* Cambridge University Press.

Jaber, A., Zhang, J., & Bareinboim, E. (2018). *Causal identification under Markov equivalence.* ArXiv. [https://doi.org/10.48550/arXiv.1812.06209](https://doi.org/10.48550/arXiv.1812.06209)

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science,* 353(6301), 790–794. [https://doi.org/10.1126/science.aaf7894](https://doi.org/10.1126/science.aaf7894)

Jerzak, C. T., Johansson, F., & Daoud, A. (2022). *Estimating causal effects under image confounding bias with an application to poverty in Africa.* ArXiv. [https://doi.org/10.48550/arXiv.2206.06410](https://doi.org/10.48550/arXiv.2206.06410)

Jerzak, C. T., Johansson, F., & Daoud, A. (in press). *Image-based treatment effect heterogeneity.* In *Proceedings of the Second Conference on Causal Learning and Reasoning, Proceedings of Machine Learning Research.*

Johansson, F. D., Sontag, D., & Ranganath, R. (2019). *Support and invertibility in domain-invariant representations.* ArXiv. [https://doi.org/10.48550/arXiv.1903.03448](https://doi.org/10.48550/arXiv.1903.03448)

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science,* 349(6245), 255–260. [https://doi.org/10.1126/science.aaa8415](https://doi.org/10.1126/science.aaa8415)

Ju, C., Wyss, R., Franklin, J. M., Schneeweiss, S., Häggström, J., & van der Laan, M. J. (2017). *Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data.* ArXiv. [https://doi.org/10.48550/arXiv.1706.10029](https://doi.org/10.48550/arXiv.1706.10029)
Kasy, M. (2018). Optimal taxation and insurance using machine learning—Sufficient statistics and beyond. *Journal of Public Economics, 167*, 205–219. https://doi.org/10.1016/j.jpubeco.2018.09.002

Keuschning, M., Lovsjö, N., & Hedström, P. (2017). Analytical sociology and computational social science. *Journal of Computational Social Science, 1*, 3–14. https://doi.org/10.1007/s42001-017-0006-5

King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press.

Kino, S., Hsu, Y.-T., Shiba, K., Chien, Y.-S., Mita, C., Kawachi, I., & Daoud, A. (2021). A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health, 15*, Article100836. https://doi.org/10.1016/j.ssmph.2021.100836

Kleinberg, J., Liang, A., & Mullainathan, S. (2017). *The theory is predictive, but is it complete? An application to human perception of randomness*. ArXiv. https://doi.org/10.48550/arXiv.1706.06974

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Knox, D., & Lucas, C. (2019). *A dynamic model of speech for the social sciences*. SSRN. https://doi.org/10.2139/ssrn.3490753

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., & Jiang, Z. (2020). Causal inference. *Engineering, 6*(3), 253–263. https://doi.org/10.1016/j.eng.2019.08.016

Kuhn, T. S. (2012). *The structure of scientific revolutions*: (50th anniversary ed.). University of Chicago Press.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2018). *Meta-learners for estimating heterogeneous treatment effects using machine learning*. ArXiv. https://doi.org/10.48550/arXiv.1706.03461

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences, 116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 4066–4076). Curran Associates. http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020).
Computational social science: Obstacles and opportunities. Science, 369(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in Medicine, 29(3), 337–346. https://doi.org/10.1002/sim.3782

Li, X., & Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. Journal of the American Statistical Association, 116(533), 353–367. https://doi.org/10.1080/01621459.2020.1844210

Liao, P., Klasnja, P., & Murphy, S. (2020). Off-policy estimation of long-term average outcomes with applications to mobile health. Journal of the American Statistical Association, 116(533), 382–391. https://doi.org/10.1080/01621459.2020.1807993

Lin, A., Merchant, A., Sarkar, S. K., & D’Amour, A. (2019). Universal causal evaluation engine: An API for empirically evaluating causal inference models. PMLR, 40, 50–58. http://proceedings.mlr.press/v104/lin19a.html

Lipton, Z. C. (2017). The mythos of model interpretability. ArXiv. https://doi.org/10.48550/arXiv.1606.03490

Lundberg, I., Johnson, R., & Stewart, B. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. SocArXiv. https://doi.org/10.31235/osf.io/ba67n

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., & Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics, 11(1), 201–234. https://doi.org/10.5194/esd-11-201-2020

Malthus, T. R. (1826). An essay on the principle of population, or a view of its past and present effects on human happiness; with an inquiry into our prospects respecting the future removal or mitigation of the evils which it occasions (6th ed.). John Murray. http://oll.libertyfund.org/title/312 on 2008-09-12

Marathe, M., & Vullikanti, A. K. S. (2013). Computational epidemiology. Communications of the ACM, 56(7), 88–96. https://doi.org/10.1145/2483852.2483871

Marini, M. M., & Singer, B. (1988). Causality in the social sciences. Sociological Methodology, 18, 347–409. https://doi.org/10.2307/271053

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.), Past, present,
and future of statistical science (pp. 537–562). Chapman and Hall/CRC.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics, 12*(2), 685–676. https://doi.org/10.1214/18-AOAS1161SF

Meng, X.-L. (2020). What is your list of 10 challenges in data science? *Harvard Data Science Review, 2*(3). https://doi.org/10.1162/99608f92.a3e88876

Merton, R. K. (1968). *Social theory and social structure*. The Free Press.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology, 45*(1), 27–45. https://doi.org/10.1146/annurev-soc-073117-041106

Molnar, C., Casalicchio, G., & Bischl, B. (2020). *Interpretable machine learning—A brief history, state-of-the-art and challenges*. ArXiv. https://doi.org/10.48550/arXiv.2010.09337

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.

Mozer, R., Miratrix, L., Kaufman, A. R., & Anastasopoulos, L. J. (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis, 28*(4), 445–468. https://doi.org/10.1017/pan.2020.1

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(2), 331–355. https://doi.org/10.1111/1467-9868.00389

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research, 49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Nie, X., & Wager, S. (2017). *Learning objectives for treatment effect estimation*. ArXiv. https://doi.org/10.48550/arXiv.1712.04912

Nie, X., & Wager, S. (2018). *Quasi-oracle estimation of heterogeneous treatment effects*. ArXiv. https://doi.org/10.48550/arXiv.1712.04912
Noble, D. (2002). The rise of computational biology. *Nature Reviews Molecular Cell Biology*, 3(6), Article 6. [https://doi.org/10.1038/nrm810](https://doi.org/10.1038/nrm810)

Okori, W., & Obua, J. (2011). Machine learning classification technique for famine prediction. *Proceedings of the World Congress on Engineering*, 2, 4–9.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60. [https://doi.org/10.1145/3241036](https://doi.org/10.1145/3241036)

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), 579–595. [https://doi.org/10.1214/14-STS486](https://doi.org/10.1214/14-STS486)

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pearson, K. (1911). *The grammar of science* (3rd ed.). Adam and Charles Black.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

Peysakhovich, A., & Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133, 373–384. [https://doi.org/10.1016/j.jebo.2016.08.017](https://doi.org/10.1016/j.jebo.2016.08.017)

Popper, K. (2002). *The logic of scientific discovery*. Routledge.

Reddy, S. G., & Daoud, A. (2020). Entitlements and capabilities. In E. C. Martinetti, S. Osmani, & M. Qizilbash (Eds.), *The Cambridge handbook of the capability approach* (pp.677–685). Cambridge University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should i trust you?”: Explaining the predictions of any classifier. ArXiv. [https://doi.org/10.48550/arXiv.1602.04938](https://doi.org/10.48550/arXiv.1602.04938)

Risi, J., Sharma, A., Shah, R., Connelly, M., & Watts, D. J. (2019). Predicting history. *Nature Human Behaviour*, 3(9), 906–912. [https://doi.org/10.1038/s41562-019-0620-8](https://doi.org/10.1038/s41562-019-0620-8)

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515), 988–1003. [https://doi.org/10.1080/01621459.2016.1141684](https://doi.org/10.1080/01621459.2016.1141684)

Robins, J. M., & Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. In C. Glymour & G. Cooper (Eds.), *Computation, Causation, and Discovery* (pp. 305–
321). AAAI Press/MIT Press.

Rosenbaum, P. R. (2020). *Design of observational studies.* Springer International Publishing.  
[https://doi.org/10.1007/978-3-030-46405-9](https://doi.org/10.1007/978-3-030-46405-9)

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence,* 1(5), Article 5. [https://doi.org/10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x)

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., & Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning,* 11(1), 1–96. [https://doi.org/10.1561/2200000070](https://doi.org/10.1561/2200000070)

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., & Vespignani, A. (2012). Digital epidemiology. *PLOS Computational Biology,* 8(7), Article e1002616. [https://doi.org/10.1371/journal.pcbi.1002616](https://doi.org/10.1371/journal.pcbi.1002616)

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age.* Princeton University Press.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., & Compton, R. J. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences,* 117(15), 8398–8403.

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the Special Collection on the Fragile Families Challenge. *Socius,* 5. [https://doi.org/10.1177/2378023119871580](https://doi.org/10.1177/2378023119871580)

Salganik, M. J., Maffeo, L., & Rudin, C. (2020). Prediction, machine learning, and individual lives: An interview with Matthew Salganik. *Harvard Data Science Review,* 2(3). [https://doi.org/10.1162/99608f92.eecdfa4e](https://doi.org/10.1162/99608f92.eecdfa4e)

Sanders, N. (2019). A balanced perspective on prediction and inference for data science in industry. *Harvard Data Science Review,* 1(1). [https://doi.org/10.1162/99608f92.644ef4a4](https://doi.org/10.1162/99608f92.644ef4a4)

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology,* 185(1), 65–73. [https://doi.org/10.1093/aje/kww165](https://doi.org/10.1093/aje/kww165)

Schultz, M. (2018). The problem of underdetermination in model selection. *Sociological Methodology,* 48(1), 52–87. [https://doi.org/10.1177/0081175018786762](https://doi.org/10.1177/0081175018786762)

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science,* 12(1), 487–508. [https://doi.org/10.1146/annurev.polisci.11.060606.135444](https://doi.org/10.1146/annurev.polisci.11.060606.135444)

Sen, A. K. (1981). *Poverty and famines: An essay on entitlement and deprivation.* Clarendon.
Shah, R. D., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics, 48*(3), 1514–1538. [https://doi.org/10.1214/19-AOS1857](https://doi.org/10.1214/19-AOS1857)

Shiba, K., Daoud, A., Hikichi, H., Yazawa, A., Aida, J., Kondo, K., & Kawachi, I. (2021). Heterogeneity in cognitive disability after a major disaster: A natural experiment study. *Science Advances, 7*(40). [https://doi.org/10.1126/sciadv.abj2610](https://doi.org/10.1126/sciadv.abj2610)

Shiba, K., Daoud, A., Hikichi, H., Yazawa, A., Aida, J., Kondo, K., & Kawachi, I. (2022). Uncovering heterogeneous associations between disaster-related trauma and subsequent functional limitations: A machine-learning approach. *American Journal of Epidemiology*, Article kwac187. [https://doi.org/10.1093/aje/kwac187](https://doi.org/10.1093/aje/kwac187)

Shiba, K., Daoud, A., Kino, S., Nishi, D., Kondo, K., & Kawachi, I. (2022). Uncovering heterogeneous associations of disaster-related traumatic experiences with subsequent mental health problems: A machine learning approach. *Psychiatry and Clinical Neurosciences, 76*(4), 97–105. [https://doi.org/10.1111/pcn.13322](https://doi.org/10.1111/pcn.13322)

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. [https://doi.org/10.1214/10-STS330](https://doi.org/10.1214/10-STS330)

Shmueli, G. (2021). Comment on Breiman’s “Two cultures” (2002): From two cultures to multicultural. *Observational Studies, 7*(1), 197–201. [https://doi.org/10.1353/obs.2021.0010](https://doi.org/10.1353/obs.2021.0010)

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354–359. [https://doi.org/10.1038/nature24270](https://doi.org/10.1038/nature24270)

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). A Bradford Book.

Sun, R. (2008). *The Cambridge handbook of computational psychology*. Cambridge University Press.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *PMLR, 70*, 3319–3328. 
[http://proceedings.mlr.press/v70/sundararajan17a.html](http://proceedings.mlr.press/v70/sundararajan17a.html)

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., & Wager, S. (2020). Policystree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software, 5*(50), Article 2232. [https://doi.org/10.21105/joss.02232](https://doi.org/10.21105/joss.02232)

Swedberg, R. (2017). Theorizing in sociological research: A new perspective, a new departure? *Annual Review of Sociology, 43*(1), 189–206. [https://doi.org/10.1146/annurev-soc-060116-053604](https://doi.org/10.1146/annurev-soc-060116-053604)

Tesfatsion, L., & Judd, K. L. (2006). *Handbook of computational economics: Agent-based computational economics*. Elsevier.
Turco, C. J., & Zuckerman, E. W. (2017). Verstehen for sociology: Comment on Watts. *American Journal of Sociology, 122*(4), 1272–1291. [https://doi.org/10.1086/690762](https://doi.org/10.1086/690762)

van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science*. Springer International Publishing. [https://doi.org/10.1007/978-3-319-65304-4](https://doi.org/10.1007/978-3-319-65304-4)

van der Laan, M. J., & Rose, S. (2011). *Targeted learning—Causal inference for observational and experimental data*. Springer.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics, 2*(1), Article 11. [https://doi.org/10.2202/1557-4679.1043](https://doi.org/10.2202/1557-4679.1043)

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Verhagen, M. D. (2022). A pragmatist’s guide to using prediction in the social sciences. *Socius, 8*. [https://doi.org/10.1177/23780231221081702](https://doi.org/10.1177/23780231221081702)

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danhelka, I., Huang, A., Sifre, L., Cai, T., Aqapiou, J. P., Jaderberg, M., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*(7782), 350–354. [https://doi.org/10.1038/s41586-019-1724-z](https://doi.org/10.1038/s41586-019-1724-z)

Watts, D. J. (2014). Common sense and sociological explanations. *American Journal of Sociology, 120*(2), 313–351. [https://doi.org/10.1086/678271](https://doi.org/10.1086/678271)

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour, 1*(1), 1–5. [https://doi.org/10.1038/s41562-016-0015](https://doi.org/10.1038/s41562-016-0015)

Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S. D., Ju, C., & Franklin, J. M. (2017). Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*. Advance online publication. [https://doi.org/10.1097/EDE.0000000000000762](https://doi.org/10.1097/EDE.0000000000000762)

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. [https://doi.org/10.1177/1745691617693393](https://doi.org/10.1177/1745691617693393)

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications, 11*(1), Article 1. [https://doi.org/10.1038/s41467-020-16185-w](https://doi.org/10.1038/s41467-020-16185-w)

Zhang, J., & Bareinboim, E. (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. *PMLR, 119*, 11012–11022. [https://proceedings.mlr.press/v119/zhang20a.html](https://proceedings.mlr.press/v119/zhang20a.html)
Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). The AI economist: Improving equality and productivity with AI-driven tax policies. ArXiv. https://doi.org/10.48550/arXiv.2004.13332

©2023 Adel Daoud and Devdatt Dubhashi. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

The preview image was created with the assistance of DALL-E 2, with the following cue: “A surreal painting by Dalí of a tree of knowledge emerging from the three statistical cultures.”

**Footnotes**

1. See Hedström and Ylikoski (2010) for an overview of various definitions of causality.
2. Usually, $\beta$ refers to a super-population parameter, but here we are using $\beta_{\text{poverty}}$ to refer to a finite-population quantity.
3. While our argument may hold for the natural sciences, our account focuses on the applied (quantitative) social sciences.
4. For causal inference, we use notations from both do-calculus (Pearl, 2009) and the potential outcomes framework (Imbens & Rubin, 2015).
5. Systems that do not map into a well-defined causal order may map to other graphical representations (e.g., sequence, equilibrium, or transition graphs).
6. As famine is binary, then we should technically be using a link function such as a logit or probit transformation.
7. https://fews.net/.
8. Undirected graphs are called Markov networks (Pearl, 2009).
9. This way of probing the model is called post hoc analysis. Other methods to aid in interpretability are called integrated gradients (Sundararajan et al., 2017) and local surrogate models (Ribeiro et al., 2016). While in the former, the gradient (derivative) of the model’s parameters space is probed, the latter approach trains a less complex model (e.g., linear model) on the predictions to probe patterns in the original model.
10. As pointed out by James Bailie, who generously commented on our manuscript, explainable and interpretable AI may themselves be two examples of the HMC, as they are not the focus of the traditional
AMC. In principle, we agree with that statement to the extent that those AI algorithms probe causal relationships. 

11. https://www.nextrembrandt.com.

12. https://openai.com/blog/musenet/.

13. The now infamous example of Google Flu is another illustration of where AMC can go wrong. The Google Flu Trends algorithm could predict flu outbreak with high accuracy, based on what people were looking for over Google Search. However, its prediction accuracy went down when the swine flu started. With that type of flu, media coverage increased to the degree that Google’s flu algorithm started to overestimate the presence of flu. Research in ML explains these types of problems as distribution shift.

14. In the notations of do-calculus we use $do(W=w)$ to capture the intervention of $W$ equal to a specific value $w$. The average causal effect is then denoted $E(Y=y|do(W=1)) - E(Y=y|do(W=0))$ (Pearl, 2009).

15. We should use a link function because $famines_i$ is a binary variable, but leave it out for sake of simplicity.

16. https://www.kaggle.com/.