[RESUBMIT] What Good is a Regression?

Inference to the Best Explanation and the Practice of Political Science Research*

Arthur Spirling† Brandon M. Stewart‡

Abstract

We argue that almost all empirical social science research should employ a mode of argumentation called "Inference to the Best Explanation" (IBE). While elements of IBE appear widely, it is seldom consciously described, leading to confusion about the role that evidence plays in social science claims. We describe what IBE is and its implications for the evaluation of studies of various types—from quantitative description, to traditional regression studies, to work with modern causal identification. We contend that debates over the merits of these approaches should be understood as debates over the relative weight scholars place on quality of explanations, the quality of evidence and the bridge between the two. Seeing the process this way helps highlight the merits of different research traditions and explains common pathologies of the same. Informed by IBE, we clarify the motivation for certain advice in regression studies and offer guidance on best practice.

*First version: March 25, 2022. This version: March 8, 2024. Amy Catalinac, Keith Dowding, Andy Eggers, Ari Hyytinen, Stephen Jesse, Robert Kubinec, Jacob Montgomery, Zagreb Mukerjee, Kevin Munger, Michael Strevens and Hannah Waight provided very helpful comments on an earlier draft. We thank audiences at MPSA and TexMeth for helpful feedback.

†Professor of Politics, Princeton University (arthur.spirling@princeton.edu)

‡Associate Professor of Sociology and Office of Population Research at Princeton University (bms4@princeton.edu)

1 Introduction

Regression analysis is now ubiquitous in political science yet debate rages over what it can show and how it should be used. Some scholars believe that without a plausible causal identification strategy, a given regression offers very little of value for most studies of politics (e.g. Samii, 2016). Others suggest that this focus is unhelpful and misleading; for them, the discipline needs a sustained move away from a tight focus on inference and towards suggesting and testing explanations (e.g. Huber, 2013). For researchers conducting and evaluating regressions themselves, confusion looms: which of these two philosophical visions is correct, and on what grounds? Our answer is that it is neither: the discipline currently presents a false choice. That is, there is no single 'right' option, and believing that there is one misunderstands how science can, does and should proceed. In fact, a single mode of inference unites and is compatible with both positions; researchers should be aware of it and describe the contributions of their work specifically in its terms. That framework is known as *abductive inference* in general, and *Inference to the Best Explanation* (IBE) in particular. It appears to be almost unknown to the discipline (though see Dowding and Miller, 2019; Tavory and Timmermans, 2014; Heckman and Singer, 2017, on prediction in political science, qualitative research design in sociology, and empirical work in economics, respectively). This is surprising and unfortunate. It is surprising, because as we will show, almost all published empirical work makes use of it.[1] That includes the types of studies both Samii and Huber, and many in-between, would like to see more of. It is unfortunate, because ignorance of IBE has led to much confusion over what we are—and should be—trying to do when running regressions.[2] Our position is that scholars should use IBE explicitly and must

[1] One possible alternate position is that IBE *is* known to scholars via its similarity to Bayesian updating (see e.g. Fairfield and Charman, 2022; Humphreys and Jacobs, 2015, for an empirical framework relying on Bayesian principles). But first, this relationship is itself debated (e.g. Douven, 2013; Henderson, 2020). And second, the merits of our arguments below do not depend on fine distinctions between these positions.

[2] We focus specifically on regression, because this is our expertise and is at the center of much recent debate. But we acknowledge that a version of what we argue applies much more generally (e.g. to qualitative

be aware of its implications when pursuing and evaluating research. To be clear: there can be value in regressions where the treatment effect is plausibly causally identified, and there can be value in regressions where it is not. More specifically, gathering evidence for a causal claim does not require estimating a causally identified parameter—these are separate but related activities that are often confused for each other. Understanding exactly where and how this idea bites requires adopting IBE.

We will make two arguments—one broad, one narrow. Our observation is that as a field, progress occurs when we narrow down the number of plausible "explanations" (a term we will discuss below) for a given phenomenon by collecting facts that either are or are not consistent with those explanations. Research contributions therefore fall into three overlapping types: the generation of explanations, the production of facts, and the updating of explanations in light of those facts. The first of these involves theorizing about social science relationships in the world: "what factors might explain why civil wars occur in some places rather than others?" is an example. The second uncovers new information about those relationships, which might involve counterfactuals or not: "do the states that have more natural resources also have more civil wars empirically?" would be a descriptive example. The third category of contribution is more ambiguous, but nonetheless common. It involves a broader critical assessment of whether and how the (new) facts imply that a given explanation is more or less plausible, given what else we know about the world. For example, a paper might theorize that natural resources are a cause of civil wars. And it might then show that civil wars are indeed empirically more common in states with more natural resources (though this is neither necessary nor sufficient for the causal claim). Given the provided set of explanations, the third type of contribution suggests what facts we might expect to be consistent or inconsistent with those explanations and how we should update our beliefs. This might involve presenting qualitative evidence for a given mechanism, or verifying that other relationships implied by

evidence).

the explanation hold in the sample or in other places; it might involve technical checks on the sensitivity of a counterfactual claim to unobserved confounding (e.g. Blackwell, 2014; Cinelli and Hazlett, 2020), or alternative measurement strategies as regards the main variables.

IBE requires each of these three pieces to function and scholars reasonably disagree on where the primary focus should be. Some prioritize new explanations, others emphasize the quality of each individual fact; still others marshal several individually less-convincing pieces of evidence together into a coherent whole. Our point is that all of these tasks have value, and that this position is completely consistent with the way that science can and should operate. Otherwise put, there is no axiomatic argument for dismissing or promoting work that does one but not others of these things. This does not mean anything goes. For example, we will make the case that we should prefer explanations that apply broadly, are transportable across settings, and that have elaborated implications. Other scholars can have different desiderata for what an explanation should be and should do (see, e.g. Ashworth, Berry and Bueno de Mesquita, 2021, on "commensurability"). In any case though, understanding that IBE is the preferred reasoning mode of social science gives a new and precise impetus to certain practices—such as explicitly embracing exploration (Munger, Guess and Hargittai, 2021) and/or avoiding HARKing (Kerr, 1998). We explain these connections below.

Our more narrow argument concerns the specific debate over the centrality of causal identification to quantitative work. This is most obviously laid out in work by Huber (2013) and Samii (2016). Huber contends that a "laser focus on causal identification" is leading political science away from answering its most interesting questions; meanwhile Samii argues that without causal inference we are in the business of the "mass production of quantitative 'pseudo-general pseudo-facts'" (941). We argue that *any* estimate, however well causally-identified, does not produce general knowledge without IBE. IBE is necessary to say how it connects to theory, where it is valid, which part of the treatment produces the effect, and so on. From there, we argue that much of the apparent 'debate' is miscast, with respondents not

disagreeing as much as focusing on different parts of the process. The goal of quantitative social science is not limited to uncovering causally identified facts; the goal should be to harness (many) pieces of evidence to obtain an inference to the best explanation—both within and across studies. Just because it is typically easier to link causal estimands—as opposed to non-causal estimands—to explanations does not make this any less true. More directly, Huber was right to be concerned about the fate of the study of politics: it cannot proceed without work that generates and evaluates explanations and it will become moribund without that element.

Before setting out the structure of our paper, we define terms. *Inference to the Best Explanation* (IBE)—a phrase coined in Harman (1965)—is a mode of inference that describes everyday reasoning. Given a set of facts, we infer that from a set of possible explanations, the one that 'best' explains the evidence is the one most likely to be true. The best explanation might be one that is simpler, more complete, or possessing some other desirable property. IBE is a form of "abductive inference" inference (sometimes "abduction"), in that it involves a non-deterministic link between observation and conclusion, via a (typically causal) story that explains the outcomes we saw. Social scientists will recognize it most clearly in work where authors lay out multiple theories (explanations) and adjudicate between with them with a variety of regression-based tests (facts/evidence). This is not just a difference of terminology: IBE emphasizes that the *candidate explanations* and the way they are consistent or inconsistent with facts have as important a role in the credibility of the inference as the evidence.[3] Thus, credible estimation of a descriptive fact or causal query is only one piece of the larger inferential framework. We use 'regression' as a catch-all for any parametric or non-parametric model in which there is an outcome (dependent variable) that is a function of at least one predictor (independent variable). Typically, we are using it to characterize a property (generally the expectation) of the conditional distribution of the outcome given

[3]See e.g. Gelman and Imbens (2013, 3) for a related discussion on when patterns "need an explanation."

the predictors; it tells us how the *prediction* will change as we vary a particular input. This includes linear regression as a special case, but also subsumes generalized linear models, and various techniques traditionally deemed part of machine learning.[4] In some circumstances, these techniques yield a causally identified quantity, but our explicit claim is that this not required for them to be useful in IBE. That is, describing an association in the world might be sufficient to update the relative plausibility of competing explanations.

We next (Section 2) sketch the ways in which regression is currently deployed in the discipline. In Section 3 we provide a definition of IBE before showing how it is used (knowingly or not) in the theory and practice of social science research. In Section 4 we explain why understanding current practice as IBE matters for how we pursue and evaluate research. We conclude in Section 5.

2 Regression in Contemporary Work

Consider a common scenario. An author provide a series of accounts of how some social or political phenomenon unfolds in the world, followed by a numbered list of hypotheses (e.g. H1a, H1b, H2...) where the alternative hypothesis (as opposed to the null hypothesis) is consistent with one of these worldviews. The author then presents a regression table and discusses the association of independent variable(s) on outcome(s) of interest. The data is observational and the result is not explicitly claimed to be causal, but the hypotheses imply certain variables are of more substantive interest than others. In addition, the author controls for (i.e. conditions on) various other variables in the regression but does not provide particularly clear assertions about the nature of the assumed confounding, any potential post-treatment bias, or the implied causal structure (in the sense of Keele, Stevenson and Elwert, 2020). Over various specifications, the relationship between the variable(s) of interest and

[4]See e.g. Aronow and Miller (2019) for a textbook account of linear regression as an approximation of an unknown conditional expectation function.

the outcome(s) are consistent, insofar as the sign and direction of the key coefficients remain similar across myriad model specifications splayed across the columns of the table. The author declares victory for their chosen explanation. Put crudely: what is this regression good for, and how might we assess its merits?

Contrast this scenario with a case where a single theory is described which implies the effect of a general type of intervention on the world. An experiment, natural experiment, or designed observational study is conducted and leads to a causally well-identified estimate of a related counterfactual in a (possibly, undefined) subpopulation. These studies might also include a regression table of some kind, typically with a well-defined treatment variable and a coefficient which is assumed to be representative of some kind of average of causal effects. The author declares victory for their chosen theory about the general benefit (or harm) of the general type of intervention. Are such approaches inherently superior to the one described above?

These abstractions are extreme but capture the spirit of many studies currently being published in top journals. We have assumed that the ultimate goal in both cases is to evaluate a particular theory or explanation of something in the world and not, e.g. prediction for its own sake (in the sense of Cranmer and Desmarais, 2017). In a thin sense, there is no dispute about what a regression does in either case: it approximates a conditional expectation function. But when, and in what ways, that conditional expectation function is useful is the subject of considerable debate.[5] It is increasingly popular to interpret the first scenario as a failed case of the latter where the causal assumptions are simply not plausible. However, we argue that this is not simply a matter of disagreement over what assumptions are plausible

[5]Claims regarding regressions in published work often revolve around the fact that a coefficient is stable across specifications. It is worth articulating what this means. Consider a binary treatment variable. The coefficient in the table is the parametric approximation of the difference in means between the treated and untreated averaged over the strata defined by the other covariates. The stability of the coefficient implies that (given the model approximation) this difference remains approximately the same as we change the subgroups over the different models in the table. This is not particularly compelling evidence either for or against the idea that this represents a causal effect.

when; it is a question of competing visions about what the work of the discipline should be and where effort is most profitably applied.

Although a relatively recent development in the history of social science, the "credibility revolution" (Angrist and Pischke, 2010) is a natural place to begin this discussion. The central idea is that making causal statements is difficult with observational data and can only be done in a more limited set of circumstances than may be initially realized. Indeed, per Samii (2016) (see also e.g. Gelman and Hill, 2006; Gerber et al., 2014; Keele, Stevenson and Elwert, 2020), regressions without thought about these issues may be actively misleading. Consequently, scholars must search for a "strong design" in order to make "persuasive" causal claims (Sekhon, 2009, 503). This is hard to achieve even in seemingly propitious circumstances where, for example, treatment and control may be randomized but the groups thus created are not comparable (e.g. Sekhon and Titiunik, 2012). While the technical claims of these studies are not in doubt, there has been much disagreement about what the credibility revolution should mean for the focus of political science research in general.

Some scholars, like Huber (2013) (see also Huber 2017, Ch 6; Clark and Golder 2015; Binder 2020), argue that the turn to causal inference is potentially troubling for two reasons. The first, is that many substantively interesting phenomena do not naturally lend themselves to such work (because e.g. the treatment cannot be plausibly randomized), and thus we see less effort to study such questions. The second reason is that focusing on identification opportunities crowds out theory development: the claim is that traditional (not plausibly causal) regression designs help us refine our understanding of relationships in observational data. In the context of randomized controlled trials (RCT), Deaton and Cartwright (2018) make an allied argument. That is, the results of (necessarily) specific RCTs cannot be easily extrapolated to broader questions of interest in a field. By contrast, scholars like Samii (2016) contend that these fears are somewhere between overblown and exactly wrong. More specifically, we should avoid using traditional designs that generate "pseudo-general pseudo-

facts" (Samii, 2016, 1). And such entities are a bad basis for either trying to understand phenomena or building theories about them. Thus, to the extent that the credibility revolution has changed practices, it has done so in a way that moves authors away from actively misleading themselves from their results. A related but distinct concern comes from those who contend that causal empiricists and formal theorists are not communicating with each other—they are "pulling apart", when they should be cooperating (Ashworth, Berry and Bueno de Mesquita, 2021).

For others, the priority is not producing causal claims (of whatever plausibility), but description. Thus we see work by Gerring (2012) that emphasizes the importance of the descriptive task as an end unto itself and independent of theory-testing. Indeed, scholars have proposed entire journals to help counter the fact that "[c]ausal research that asks the question *why* has largely taken the place of descriptive research that asks the question *what*" (Munger, Guess and Hargittai, 2021, 3, emphasis as original). Here then, regressions are informative about the state of the world in terms of associations—nothing more and nothing less. Partly in an attempt to connect this associational logic to the goal of inference, researchers have recently argued that flexible machine learning approaches—capable of including non-linear interactions—ought to be more broadly deployed for political science tasks (e.g. Montgomery and Olivella, 2018). Whatever the estimation approach, the associations are conditional on many variables. But this can make interpreting them—in terms of an all-else equal logic—difficult (Ashworth, Berry and Bueno de Mesquita, 2021). What is being described can turn out to be misleading.

Regardless of the purpose of the regressions, there has been increasing agreement on what their properties ought to be. In particular, the importance of replication and robustness in results. At one level, the concerns regard the potentially malign motivations of researchers to "p-hack" or else leave insignificant results in the "file-drawer" (e.g. Franco, Malhotra and Simonovits, 2014); for others, there are broader issues of "forking paths" wherein re-

searchers make ad-hoc but crucial decisions about data and estimation Gelman and Loken (2014). Scholars have proposed various solutions, from "multiverse analysis" of all possible choices (Steegen et al., 2016) to more focused efforts at assessing sensitivity (e.g. Imai and Yamamoto, 2010; Blackwell, 2014; Cinelli and Hazlett, 2020). Related in spirit, but different in practice, other authors have suggested methods for incorporating distributional assumptions about bias (as in, the difference between an estimated coefficient and the 'true' causal effect) into more nuanced interpretations of regression results (Little and Pepinsky, 2021).

Of course, the logic thus far assumes that authors are sufficiently clear about what they are estimating to effectively connect the results to explanations or broader theory in a credible way. Lundberg, Johnson and Stewart (2021) point out that this is often not the case. This leads to situations where debates can be entirely about disconnects over the target estimand. Lieberson and Horwich (2008) worries that the link between theory and evidence has so substantially frayed that social science is merely 'mimicking' science. While Lundberg, Johnson and Stewart (2021) focus on clarity about what estimands a researcher is targeting (and how unobservable estimands are connected to observable data), our focus is on how the chosen estimands (whether descriptive or causal) are marshalled to make claims about the world.

The central challenge is that authors rarely explicitly state their philosophical underpinnings. There are clearly instinctual understandings of what makes their arguments compelling, but this is different from having an explicit, common framework for assessing diverse evidence. We claim that such a framework already exists for accommodating all of these positions, and that it is Inference to the Best Explanation. It should be explicitly acknowledged. We now define IBE, before making this point with reference to studies in the field.

3 What is *Inference to the Best Explanation*?

Abductive inference, in the form of IBE, is ubiquitous in scientific enquiry (e.g. Harman, 1965; Boyd, 1984; Douven, 1999; Lipton, 2003). The study of politics is no exception (Dowding and Miller, 2019). Introductory accounts (e.g. Psillos, 2002; Douven, 2021) will typically give a definition of IBE along the following lines, and with which we concur:

> Given some data D (some observations, or facts about the world), and some candidate explanations or hypotheses $E_1, \ldots, E_n$ that potentially explain D, the one that is most compatible with D is most likely to be true.

Such accounts often then discuss the use of IBE in a "classic" case, such as medicine (the discussion of the Semmelweis case in Lipton, 2003, is in this vein). Generally, the moving parts are

1. theorizing or generating candidate explanations based on an initial observation (e.g. creating a differential diagnosis on the basis of the presenting symptoms of a patient)

2. collecting facts that are relevant to implications of those explanations (e.g. ordering lab tests or checking other symptoms)

3. interpreting and discriminating between explanations in light of the newly collected facts (e.g. ruling out diseases inconsistent with the tests and/or making a final diagnosis)

The order is important: collected facts in Step 2 are helpful when they assess implications of the explanations and produce new information beyond the initial observations in Step 1.[6] Explanations can only be discriminated in Step 3, when we have collected facts well

[6]Most patterns of facts are consistent with many plausible explanations. This increases the need to have tests based on a set of facts collected after the explanations have been generated.

in Step 2. The process is also iterative, with the result of Step 3 establishing a new initial presentation that might in turn generate new explanations.

We now elucidate the moving parts of IBE—that what is meant by explanation and inference before moving to how IBE manifests in social science in practice.

3.1 Explanation

Clearly IBE relies on some definition of "explanation"—a focus for philosophers of science for some time (see e.g. Dowding, 2015, for an overview). We agree with Clarke and Primo (2012) that, in social science at least, there are broadly two reasonable understandings of this term.

First, there is an understanding that arises from the positivist tradition of Hempel (1966). In the original formulation, the explanation followed as a deduction—an "if-then" arrangement—though Hempel and others subsequently offered weaker versions of this logic. This use of explanation requires that one identify general or "covering" laws: that is, the idea that the specific case (the country, the person, the leader) under study is an instance of a broader type of unit which displays a known regularity of behavior. For example, a covering law might be a version of *Modernization Theory* (e.g. Lipset, 1959) in which societies with more assertive middle classes typically transition from feudal arrangements to democratic ones. This general law is combined with a statement about the condition of the entity of interest (collectively, the explanans). For example, the statement might be that a particular society—say, Britain in the 1820s—has an increasingly large bourgeoise. The law and the conditions of the case then produce inferences about why a specific event occurred (the explanandam). Thus our explanation of why suffrage was widened in the Great Reform Act is that Britain had the right class conditions for the particular variant of Modernization Theory to produce democracy.

The second use of "explanation" is the causal mechanical tradition arising from the work

of Salmon (1971) and others. Here, an explanation is about showing how an event occurred in terms of the causal mechanisms involved in producing it. Thus explaining why richer nations tend to be democracies might involve a (theoretical) model in which increasingly prosperous agents threaten to revolt, and elites buy them off with voting rights. Notice that one does not need a covering law: instead, the explanation connects causes with outcomes in terms of processes of influence. The way that the mechanism is described may be formal or not, and may draw on various different logics of behavior—from rational choice to psychologically models of reasoning. Indeed, there is no particular requirement that every step of the pathway between parameters be observable at all. A special case of this logic that has proved popular can be seen in Woodward (2005). There, explanations are about *counterfactuals*. That is, a treatment or action is an explanation to the extent that the (potential) outcome would be different in the absence of a given intervention (say, a larger middle class), holding all other variables constant.

Empirically, we would argue that the latter version of "explanation" has become dominant—relative to the positivist one. Indeed, starting at least with King, Keohane and Verba (1994), some have argued that explanation *must* be causal in nature, though not all scholars agree. Importantly, in agreement with Clarke and Primo (2012) and Dowding (2015), in what follows we maintain that showing a causal relationship between some treatment and outcome is neither a necessary nor sufficient condition for supporting a causal explanation. Of course, the logic of causal identification may well be *helpful* for explanation (see, e.g. Ashworth, Berry and Bueno de Mesquita, 2021, Ch 2.3 for elucidation of this point for the case of women's electoral fortunes). But it is not required, and may not be enough. In particular, we have in mind pragmatic accounts like that of Achinstein (1983), where explanations need not be causal at all——instead, they must achieve 'understanding' for a particular question as defined by a particular audience (here, political scientists).

3.2 Inference

It is straightforward to give examples of how the IBE process might work in principle. In keeping with our case above, suppose we observe that as countries become more developed (say, in per capita income terms) they are generally more likely to become democracies. One explanation might be that of *Modernization Theory*. An alternative explanation might follow this basic logic but specify that middle class activism is a product of particular social relations some centuries before (Moore et al., 1993). Other explanations might focus on the role of income inequality (e.g. Boix and Stokes, 2003) or elite responses to the threat of revolution (Acemoglu and Robinson, 2001). Abstracting from this specific example, this is a near-universal undertaking: scholars observe (essentially) the same data and attempt to provide a 'best' explanation for this data. And when they do this, they are doing IBE. We can push this point further. When scholars gather *new* data and suggest a 'best' explanation for those observations—relative to other explanations or even simply a null hypothesis—they are also doing IBE. Consequently in empirical social science, almost everyone, all of the time, is doing IBE. We will expand on this idea below, but before doing so we clarify the position and nature of IBE more broadly.

First, abductive inference is in contrast to both *deduction* and *induction*. Deduction requires that our claim must follow from our premises. A familiar case is formal theory, where we specify predicates (say, assumptions of a theoretical model) and agree on what operations one can undertake on those predicates (say, what makes for an 'equilibrium' in a given game). Logic of this kind extends to the "deductive approach" to causal inference. There we make assumptions about the data generating process, which we specify in an identification strategy. For example, we might assume that in elections, candidates that narrowly win are in all relevant ways identical to candidates that narrowly lose. Thus we can regard the treatment of winning 'as if' randomly assigned to that set of politicians (e.g. Lee, 2008). To the extent that districts that parties narrowly win subsequently run up bigger

party vote shares than places they narrowly lose, we say that this is due to the causal effect of incumbency. Here the claim of causal effect follows directly from assumptions we made about how the world works. Note that this does not require that the treated group (winners) must, with certainty, have a larger average vote share than the control group (losers); rather it is that any (statistically significant) difference in the averages is the causal effect of the difference in victor status.

In contrast, *induction* does not involve these sorts of necessary conditions. This is the case even though we may accept the truth of the premise. In line with our comparative politics example above, an inductive inference might be that a randomly chosen rich country is very likely to be democracy. Unlike for deduction though, we make no claim that this *must* follow from some identification strategy logic. In addition, and crucially unlike in abductive inference, induction does not require we offer a causal 'story' as to *why* we expect a rich country to be democratic. Induction can simply assert that these features generally co-occur. IBE requires the extra step.

Second, however common, IBE is not (claimed to be) a perfect strategy for inference. IBE contains a logical fallacy: "affirming the consequent." From our case above, for example, if *Modernization Theory* is correct, then it follows that we would see a particular pattern of democratization. But seeing that particular pattern cannot be conclusive evidence that *Modernization Theory* is correct (Clarke and Primo, 2012, make a similar point in their discussion of 'models'). Second, we have no guarantees that the set of explanations from which we are purportedly selecting the 'best' one contains the truth. Indeed, the fact that political scientists continue to propose new explanations for the development data we observe suggests that the field as a whole has not yet reached the end point of this search. Of course, one does not need to believe that a given inference method is perfect (or even coherent—see, e.g., Van Fraassen, 1989; cf Douven, 1999) for it to be popular in practice. And if it is popular, it is important to understand its characteristics and implications.

In quantitative social science work, it is often the case that only one explanation is offered, and that it is tested against *only* a null hypothesis (Gross, 2015). Of course, the null hypothesis is not itself an explanation for the data (see McShane et al., 2019), so failing to reject the null is potentially awkward in the world of IBE. That is, we have not found evidence consistent with our preferred explanation, but nor have we made an inference to another, 'best', explanation. This does not mean that the study is of no value. Obviously, as with medicine, it can be important in an individual case to show that a particular explanation is inconsistent with the available facts (e.g. we find that a patient tests negative for a shrimp allergy, and thus might likely be able to eat shrimp). More optimistically, at a *field* level of surveying many studies regarding similar data, we can presumably claim that a process consistent with IBE is taking place. That is, when we look at a literature at a high level, we are collectively generating new explanations, and making an inference to the best one by rejecting those that are not supported. But some care is required in this weaker, aggregated understanding of IBE: simply generating a large number of bad (unsupported) explanations is not especially helpful, nor is generating a lot of facts unrelated to any explanations of interest.

3.3 IBE in Social Science: Practice

One part of our argument is that many scholars are already doing things compatible with IBE, whether they know it or not. Our claim is that they should embrace IBE explicitly and think carefully about what the implications of the framework for their work. We now show how different types of social science work could and should proceed on this score.

3.3.1 IBE in Observational Studies

In one of the most highly cited articles ever published in the *American Political Science Review*, Fearon and Laitin (2003) seek to explain why the 20th Century saw a notable rise

in civil conflict. They pose the question: "What explains the recent prevalence of violent civil conflict around the world?" (75). They then enumerate a set of candidate explanations from conventional wisdom as to what makes countries susceptible to civil war: the end of the Cold War and associated changes in the international system, ethnic or religious diversity, and ethnic or broad political grievances. Their fourth (preferred) explanation is conditions that favor insurgency, including weak central governments, positive shocks to insurgent capabilities, and rough terrain. Much of the article is devoted to detailing 10 different empirical regularities (framed as hypothesis tests) that would be implied by different explanations for civil wars. While the explanations themselves are essentially causal in nature, the tests are mostly descriptive or predictive in nature. Fearon and Laitin (2003) describe the conditions that they would expect to see in the world were each explanation best. They then test these conditions using five different regression models—containing thirteen predictors—on cross-country data.

Fearon and Laitin (2003) has been the subject of critiques on methodological grounds, many of which have emphasized the concerns around post-treatment bias (e.g. Acharya, Blackwell and Sen, 2016) and thus challenged the empirical credibility of the tests (e.g. Samii, 2016). Focusing on claims about the relationship between economic shocks and violence, Ashworth, Berry and Bueno de Mesquita (2021, Ch 9.2) note that the entangled nature of the mechanisms concerned make it hard to know what association we would expect to see between economic performance and civil conflict even if the economy was driving the conflict. This is one of the challenges of IBE: even when trying to apply descriptive or predictive tests, causal reasoning is generally necessary to determine what we would expect to see *descriptively* given that the explanation were true. Of course, some explanations can be more easily removed from consideration: Fearon and Laitin (2003) quickly dismiss the first common wisdom explanation—the increase is due to the changes at the end of the Cold War—by showing in their first figure that civil war had been steadily rising since at least

1950.

Regardless of whether one finds the tests convincing, the strategy here is one of inference to the best explanation. Thus we can evaluate the contribution of the work not just on the quality of empirical evidence, but also on the development of the candidate explanations and the connection of those explanations to the tests. Even if we believe the tests of the ten hypotheses are convincing, this does not rule out other explanations they do not consider, it simply suggests that their preferred explanation is the best of the four they offered. This makes the credibility of their conclusion—and the policy implications they draw from it at the end of the article—turn as heavily on this candidate set of explanations as the empirical credibility of the tests.

3.3.2 IBE in Experiments

While IBE helps to reinterpret what is happening in observational studies, it is also the mode of inference used in randomized experiments that target specific causal quantities. This is most obvious in the context of laboratory experiments that capture an artificial form of behavior that the authors argue generalizes to a broader class of real-world settings. This process of connecting the lab experiment to real-world phenomena of interest is one of IBE. For example. in their paper "Winners don't punish", Dreber et al. (2008) aim to assess claims about the role of self-sacrificing punishment in maintaining cooperation. They bring subjects into the lab to play a form of repeated Prisoner's Dilemma games where—in addition to the usual cooperate and defect—there was a third option to "punish" by paying 1 unit of money to have the other player lose 4 units. They find that those who perform well don't use the ability to exact the punishment and that "this suggests that costly punishment behaviour is maladaptive in cooperation games and might have evolved for other reasons" (Dreber et al., 2008, 1). Presumably, no reader of the article primarily cares about the capacity of individuals to play a Prisoner's Dilemma game in a lab. The implicit argument

is that there is a common set of behaviors that govern both how humans play artificial games in the lab and how they cooperate in higher-stakes real-world settings. The empirical fact—the counterfactual involving an option to 'punish'—is well-established, but the inference to a broader explanation (or set of explanations) is limited by the artificial construction of the experiment.

The logic of IBE also pertains to field experiments which more narrowly tailor their interventions to real-world scenarios. Consider for example, the audit study in Bertrand and Mullainathan (2004) for examining discrimination. The motivating observation is that, in the United States, Black prospective workers are twice as likely to be unemployed as white prospective workers. This is consistent with a range of explanations including racial discrimination at the point of application or differences in the supply of applicants (presumptively due to prior history of racial discrimination). The authors design a method of data collection that adjudicates these explanations: they submitted the same resumes to different jobs, swapping out the names with a set intended to convey that the applicant was white or Black. The data is that a list of (protypically) white names received 50% more callbacks than a list of (protypically) Black names. Since the resumes were the same and we might reasonably believe that employers do not have especially strong preferences towards names *per se*, they infer that the best explanation is discrimination on race at the point of application.

Note that even this claim involves an inference that the driving feature is signaling race and not socio-economic status or some other property of the applicant. Bertrand and Mullainathan (2004) attempt to adjudicate among these explanations by showing that the gap between the names is no higher for jobs that rely on soft skills or interpersonal interactions. They also show that their evidence is not well explained by either of the major economic theories usually used to explain discrimination. They suggest then that an alternative model based on a heuristic screening might be a better explanation for the behavior they observe.

Experiments are attractive because the randomization makes the correspondence be-

tween the observed association and the causal estimand particularly plausible. However, this 'merely' establishes a particularly reliable fact about the world; it does not directly evaluate a theory. IBE is the framework that moves us from the particular facts we observed to the explanation. Even in the case of experiments, the set of possible candidate explanations plays a crucial role in our understanding of what is a reasonable inference.

3.3.3 IBE in Policy Evaluation

A third and rapidly growing area of social science research is *policy evaluation* (see, e.g., Athey and Imbens, 2017, for discussion). Here we mean situations where the question is whether a particular intervention had a causal effect or not, which might also include attention to the nature of the effect, its magnitude and portability to other settings. The scenario could be experimental (say a training scheme randomized within a given firm) or observational (for instance, a government enacts a legislative change that affects some citizens born before a certain date but not those born after). The immediate purpose of such studies is typically practical insofar as the aim is to inform institutional action given a set of welfare goals; it is less concerned with answering a debated question in the literature. Put simply: should the goal of such studies also be IBE? Our answer is yes.

There are two parts to this position. First, though perhaps not explicit, IBE is involved in the motivation for a particular policy evaluation. That is, the researcher must have some explanation in mind for why the intervention (say, a new law about postal voting) could or should affect a given outcome (say, turnout), even if they are *a priori* unclear on its sign or magnitude. If they do not, then it is unclear why the study is taking place at all, or what we expect to learn from it. In this sense, IBE forces us to answer the question *What explanation is being assessed by this policy evaluation?*

Second, more narrowly: if a treatment effect turns out to exist, one updates about the *relative* merits of a proposed (perhaps implicit) explanation for a more general phenomenon.

Suppose—as was in fact the case—that some local councils in England introduced a 'four-day week' in which employees received the same total compensation and required tasks, but were expected to do 20% fewer hours. Suppose this improved average employee contentment and reduced costs. We would perhaps update that giving employees less work time such that they had to work more intensely is generally preferable to more time where they could work less intensely. In turn, this suggests that more traditional, Weberian explanations of institutional efficiency—suggesting the need for highly structured organizations and inflexible contracts—are more dubious than we previously thought. That is, IBE forces us to answer the question *What did you learn from this study that might apply to other cases?*

In both cases here, the key is to think beyond the specific studies themselves. That is, IBE can and should happen at a broader, more aggregate discipline level as other researchers read the results and assess what they mean for the plausibility of various explanations across contexts. Nonetheless, failure to consider IBE in a given study weakens the work because readers find it harder to know how they should be updating, and about what.

4 What IBE Means for Applied Research

To reiterate: many scholars are doing things consistent with IBE much of the time. But they should take it more seriously and consider the implications explicitly. We now turn to the practical implications of this claim. We emphasize three ideas: first, IBE often involves the bringing together of multiple pieces of evidence: e.g. many different regressions in various places, in addition to a qualitative case study on a potential mechanism. Second, explanation is obviously vital to the practice of IBE, and should be taken as seriously as inference is. Third, that exploratory analysis, because it encourages the development of explanations, is similarly crucial and currently undervalued.

4.1 Imperfect Tests Are Compatible with IBE

IBE suggests that we want to find pieces of evidence that discriminate between competing explanations of the underlying phenomenon. In practice, most pieces of evidence are consistent with more than one possible explanation (even in causally identified experiments as we demonstrated above). But this frees us from the need to search for a single glorious test that will uniquely demonstrate that our favored explanation is true—likely no such thing exists. Instead, we simply need to rule out other competing possibilities. Thus work done in the framework of IBE will often involve cobbling together imperfect evidence that jointly make a compelling case. An immediate consequence is Implication 1.

Implication 1 *The "Causal Two-Step" is not compatible with IBE. Researchers should clearly state that their explanations or estimands are causal, if this is what they are intended to be.*

The "Causal Two-Step" arises when a researcher includes controls *as if* to make the estimated effect of the treatment more causally plausible, yet *simultaneously* claims that the data does not allow for a causal interpretation of coefficients. This is not IBE. What is compatible with IBE is researchers either making the causal claim they want to make and acknowledging the possible imperfection of any one piece of evidence (regression, case study, interview) or being clear how the (non-causal) descriptive evidence helps to adjudicate between competing (causal) explanations of the world. That is, a causal identified regression coefficient is only one type of evidence that can be consistent or inconsistent with an explanation. The challenge of the large regression table full of controls which is 'non-causal' is that these 'descriptions' are extremely hard to interpret outside the context of a causal claim. As Ashworth, Berry and Bueno de Mesquita (2021) emphasize, the 'all-else-equal' logic of a regression coefficient is profoundly difficult to reason about descriptively when the conditioning set is large.

4.2 Explanation is as Central as Inference

Our contention is that the purported debate—between those who embrace (only) causally plausible designs and those who eschew them as unhelpful—is confused. Crudely: everyone should do IBE, though their focus may on different parts of it. If the analyst takes the position that what matters (most) is causal identification over generating explanations, they are focused on the *quality of inference* in IBE. That is, whatever the—potentially very narrow—set of explanations, they want to believe that the inference they make is the correct one. By contrast, scholars like Huber are emphasizing the importance of *quality of explanation* in IBE. That is, they want the data and models to tell us about the relative plausibility of many explanations, or actively develop new explanations. And they want to do this even if that same machinery cannot do a particularly good job of determining which one of those explanations is 'best'. Notice that this is *not* a question of "external" versus "internal" validity (in the sense of, e.g. Morton and Williams, 2010): the problem is not that one is sacrificing generalizability for (local) credible estimation of causal effects. The issue instead is the preferred tradeoff between inference and building explanations. This leads to Implication 2.

Implication 2 *A study can focus on the "I" aspect of IBE, or the "E" element, or the link between the two. Whether a study is good or bad is not determined by that choice.*

If the primary concern is tests of existing implications of existing theory, then quality of inference must be the priority. But this will not help with theoretical progress directly, because it is trivially true that an infinite number of theories is consistent with a particular (well-identified) causal effect. Similarly, elaborating implications of a series of explanations is only helpful insofar as those implications are sharply defined and effectively discriminate across theories. Far too often in social science, the target estimand is not specified well enough to know what the evidence is, much less whether it adjudicates well between theories

(Lundberg, Johnson and Stewart, 2021).

A more positive expression of this statement is that one can advance science in many different ways even without a specific well-identified causal fact. Studies can be valuable for offering new explanations, elaborating new implications of those explanations, or providing direct tests of the implications. All three pieces of that process need not be equally credible for a substantial contribution to collective knowledge. So if one wishes to critique a study, being as explicit as possible about where the area of concern is, and how, leads to more useful dialogue and a more actionable path forward. Implication 3 follows.

Implication 3 *Evaluate studies in explicitly IBE terms. A lack of a causally identified parameter is not a problem per se, so long as a study provides evidence that helps to adjudicate between explanations. Those explanations should be 'strong' (plausible), and all hypotheses—even those which are rejected—should map to compelling, elaborated explanations of the phenomenon under study*

In the IBE framework the value of testing hypotheses is to assess whether the evidence in the world is consistent with a particular set of explanations. If the hypothesis being tested are compelling, then updating our belief about the *absence of evidence* for an explanation is useful in addition to updating about the *presence of evidence*. For example, it is helpful to know that an explanation previously thought to be "best" is in fact weaker than believed. Or that none of the usual explanations for this sort of case (say, a rich country that is not a democracy) work well. Specifically, this is knowledge of value to a field *as a whole*—beyond a particular study. And it is at that more aggregated level that developing explanations and assessing the evidence for them should take place. This yields Implication 4.

Implication 4 *IBE is the process of ruling out alternative explanations. Report null or negative results even for preferred explanations, so that the community understands when there is an absence of evidence or conflicting evidence.*

Of course, researchers have long noted that null results are underreported (e.g. Franco, Malhotra and Simonovits, 2014). When this is combined with author and journal selection effects such as the "file drawer problem" (in the sense of Rosenthal, 1979) publication bias may be inevitable (Dickersin, 1990). Our point here is that IBE—the primary mode of inference in social science—gives a new and specific impetus to improving practice on this matter.

4.3 Exploration Matters

Explanations are developed from exploration. In an instructive analogy, Tukey (1977, 1) notes that the data analyst should be a "detective" attempting to fact find, before—in a separate stage–the jury system makes a ruling. This judicial role is where "Confirmatory Data Analysis" (CDA) steps in. Since CDA involves hypothesis *testing*, it is different to exploratory data analysis (EDA), which involves

> a focus on tentative model building and *hypothesis generation* in an iterative process of model specification, residual analysis, and model respecification

in the words of Behrens (1997, 132, emphasis added).

But in much modern political science work, the EDA (hypothesis generation) is not kept separate to CDA (hypothesis testing). The result is that researchers often try to do both and end up doing neither. Thus we see statements of hypotheses in papers that do not comport with either Fisher or Neyman-Pearson decision theory. There is no explicit null hypothesis, nor is there discussion of a test statistic; nor ultimately, is there a mutually exclusive decision to reject or fail to reject the null.[7] Instead, scholars write of statistical evidence "partially consistent" with a hypothesis (Thompson, 1975, 474); or they talk of hypotheses which are

[7]See Mayo (1996, Ch 12) for a nuanced discussion on the historical relationship between Peirce's work and Neyman-Pearson statistics.

"strongly supported" by the data (Lau, 1985, 130).[8] This gives rise to Implication 5.

Implication 5 *Exploratory Data Analysis (generating explanations) should be valued per se and kept separate from Confirmatory Data Analysis (testing explanations). Be explicit when doing exploration.*

We are not the first to link the idea of EDA to Peirce's original (1878) work on abductive inference. Behrens and Yu (2003, 40), for example, note that after EDA we should " "abduct" only those [hypotheses] that are more plausible for subsequent confirmatory experimentation." To be explicit here, and in keeping with our comments above, we contend that among a set of otherwise similar explorations, a given exploration is most valuable when it gives rise to testable, elaborated and general explanations. For example, Wikipedia editors are disproportionately male, relative to the world population (Ford and Wajcman, 2017)—they are presumably also disproportionately taller than the median height. But one of these explorations (and descriptions) is more interesting and more general than the other for most scholars of social science and their causal claims. Though not motivated by IBE itself, one tool that may encourage more focused explorations is pre-analysis plans; these limit the extent to which researchers can add explanations after performing the (confirmatory) analysis stage (itself coming after the EDA).

By valuing exploration, and encouraging the separation of EDA from CDA, IBE deincentivizes other problematic practices. These include Hypothesizing After the Results are Known (HARKing), defined as "presenting a post hoc hypothesis in the introduction of a research report *as if* it were an a priori hypothesis" (emphasis added) (Kerr, 1998, 197). The idea is that authors run several (perhaps many) different analyses involving statistical tests, and then write hypotheses consistent with the (strongest) results they find. Given that hypotheses should be first derived from theory, and thus offer a way to test the implications

[8] We use older studies here, because our goal is not to critique specific scholars, but rather to give examples of the types of language that can be found in many modern works.

of that theory, that HARKing reverses this basic process is immediately concerning. And indeed, the problems with this kind of interpretative overfitting have become well known. They include specific issues, such as a general tendency for Type-I errors ("false positives") to be presented as findings (Ioannidis, 2005). Related, researchers may overfit in a purely statistical sense by "p-hacking"—trying different model and data specifications in a deliberate attempt to reach a particular level of significance (Simmons, Nelson and Simonsohn, 2011). Or they might do such things less consciously, but ultimately resulting in the same problem of findings that cannot be replicated (Gelman and Loken, 2014).

HARKing occurs because scholars explore the data and then pick a particular explanation to (superficially) "test" via a hypothesis. But under IBE, this subterfuge is pointless. If analysts want to explore data to suggest new explanations, they should just do this without artificially stating hypotheses *as if* testing a pre-existing theory. And, for all the usual reasons, those explanations should not be tested with the data used to generate them. But that is a separate matter. Second, IBE suggests that description is *per se* important. In the medical example we gave above, the physician needed symptom data to make a diagnosis. So, anything that devalues "mere" description—and the incentives behind HARKing certainly do—is anathematic to IBE and the scientific approach that embodies.

A special case of these problems—and a solution that is provided by a proper consideration of IBE—is presented by much "text as data" work. There, scholars will use unsupervised techniques—including topic models (Quinn et al., 2010; Roberts et al., 2014, e.g.)—designed for summarizing and organizing a corpus. Yet researchers routinely use the output of these models to make statements about the plausibility of various theories. The issue here is not simply about the availability of forking paths arising from decisions one must make in fitting the models (see, e.g., Denny and Spirling, 2018). It is that, fundamentally, unsupervised techniques are primarily methods of discovery which can only be repurposed as tools of measurement with substantial care and validation (Grimmer and Stewart, 2013; Grim-

mer, Roberts and Stewart, 2022). The danger is what we term "PEACHing": [P]resenting [E]xplorations [A]s [C]onfirmable [H]ypotheses. This is, in a sense, the opposite of HARKing. In HARKing, techniques (like regression) that allow for hypothesis tests are used to assess a large number of possible hypotheses, and those with suitable p-values presented as if *a priori* theorized. In PEACHing, techniques that do *not* naturally facilitate hypotheses tests are used to suggest hypotheses that cannot be "tested" at all, at least on that data and with that approach (though see, e.g., Egami et al., 2022, on testing with an explicit heldout sample). What IBE makes clear is that these approaches can be used to help suggest new explanations, but the role of explanation generation must be carefully separated from the testing of those explanations. These insights yield Implication 6.

Implication 6 *Post hoc hypotheses do not provide the same inferential value because the explanation is built to fit the available facts and thus cannot be effectively distinguished from other explanations with that same set of facts. Do not engage in HARKing or PEACHing.*

5 Discussion

The credibility revolution has been a "catastrophic success"; it has so profoundly changed the practice of social science that it has created new problems in its wake. Scholars have claimed that it is crowding out questions, and approaches to those questions, that also deserve attention. Consequently, the argument goes, the discipline is all the poorer. We agree that we see more attention to ensuring claims of causal effects are plausible, but this isn't the full story. Our argument was that improving explanations, deriving implications of those explanations, and testing those implications all have a natural role in political science. A given researcher or research agenda may value one aspect over another, but they need not be in direct competition—science is ideally *cumulative* across studies with different

strengths. The abductive inference mode in which these elements simultaneously exist—known as *Inference to the Best Explanation*—is ubiquitous and long-standing. We argued that this was true of essentially all empirical set ups: from observational regressions without a plausibly identified causal effect, to randomized lab experiments where it is generally agreed that a treatment effect of some kind has been identified.

Understanding that almost all such research is or should be some variant of IBE allows for a common framework of discussion across the discipline. But it does more than that. By clarifying the importance of explanations—their development and their plausibility—IBE simplifies a series of debates in political science. We argued that if explanation matters, then exploring and describing data becomes *per se* valuable. And if that is true, there is no reason to misrepresent ('causal') hypotheses as if they occurred to the reader after running the relevant regression: the incentive to "p-hack", or to hide null results away, is reduced. More generally, an implication of recognizing the centrality of IBE means avoiding unhelpful and uncharitable comments about whether a given research question is "big" or "small"—and thus worthy of answering (or not), and with different degrees of precision. Our position is that these claims are in fact mostly displays of researchers' preferences—over what they value in terms of the components of the IBE framework. If these preferences are made explicit, progress is more likely.

Of course, embracing IBE is not a panacea. For example, even if we encourage exploration, researchers still have incentives to only present or assess certain types of explanations for value-based or cognitive reasons. That is, they remain "attorneys" for particular positions and it hard to correct that behavior.[9] In addition, IBE itself is notoriously silent about what makes for an ideal explanation. Fortunately, works like Ashworth, Berry and Bueno de Mesquita (2021) provide clear visions. There, the goal is to explain how one optimally arranges a new research design—rather than our broader question about how to assess other

[9]We are grateful to an anonymous referee for this point.

work in the field. Those authors suggest that what should matter is "commensurability" between the formal theory and the empirical estimation. This is a fine desiderata, though it leaves unclear how one should place relative value on different designs that lack 'perfect' commensurability. And of course, there are many social science areas—like the psychology of disgust (e.g. Schnall et al., 2008)—where one has no formal or strategic model of behavior. Finally, one may value other things in an explanation, like parsimony. It is good to be explicit about what a good explanation looks like in a particular discipline; we think embracing IBE encourages that.

All told, the answer to "what good is a regression?" depends on what you want from it. There are reasonable answers that arise from the same philosophical framework being used (implicitly) by those that particularly promote causal inferences. This does not mean anything goes. Authors should be clear about their explanations, what those explanations imply about the world and what their evidence actually tells us about those implications. The pitfall to avoid is the misrepresentation of explanation or evidence. Our view is that the standard for an empirical paper should be whether an interesting argument is well-advanced. We should evaluate arguments in the light of IBE. This still leaves open important questions such as the difficulty of making arguments with only descriptive evidence or the optimal aggregate balance of explanation and fact production in the discipline. We leave such tasks for future work.

References

Acemoglu, Daron and James A Robinson. 2001. "A theory of political transitions." *American Economic Review* 91(4):938–963.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3):512–529.

Achinstein, Peter. 1983. *The nature of explanation*. Oxford University Press on Demand.

Angrist, Joshua D and Jörn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24(2):3–30.

Aronow, Peter M and Benjamin T Miller. 2019. *Foundations of agnostic statistics*. Cambridge University Press.

Ashworth, Scott, Christopher R Berry and Ethan Bueno de Mesquita. 2021. *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press.

Athey, Susan and Guido W Imbens. 2017. "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic perspectives* 31(2):3–32.

Behrens, John T. 1997. "Principles and procedures of exploratory data analysis." *Psychological Methods* 2(2):131.

Behrens, John T and Chong-ho Yu. 2003. "Exploratory data analysis." *Handbook of psychology* 2:33–64.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review* 94(4):991–1013.

Binder, Sarah. 2020. "How we (should?) study Congress and history." *Public Choice* 185(3):415–427.

Blackwell, Matthew. 2014. "A selection bias approach to sensitivity analysis for causal effects." *Political Analysis* 22(2):169–182.

Boix, Carles and Susan C Stokes. 2003. "Endogenous democratization." *World politics* 55(4):517–549.

Boyd, Richard N. 1984. 3. The Current Status of Scientific Realism. In *Scientific realism*. University of California Press pp. 41–82.

Cinelli, Carlos and Chad Hazlett. 2020. "Making sense of sensitivity: Extending omitted variable bias." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(1):39–67.

Clark, William Roberts and Matt Golder. 2015. "Big data, causal inference, and formal theory: Contradictory trends in political science?: Introduction." *PS: Political Science & Politics* 48(1):65–70.

Clarke, Kevin A and David M Primo. 2012. *A model discipline: Political science and the logic of representations*. Oxford University Press.

Cranmer, Skyler J and Bruce A Desmarais. 2017. "What can we learn from predictive modeling?" *Political Analysis* 25(2):145–166.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210:2–21.

Denny, Matthew J and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2):168–189.

Dickersin, Kay. 1990. "The existence of publication bias and risk factors for its occurrence." *Jama* 263(10):1385–1389.

Douven, Igor. 1999. "Inference to the best explanation made coherent." *Philosophy of Science* 66:S424–S435.

Douven, Igor. 2013. "Inference to the best explanation, Dutch books, and inaccuracy minimisation." *The Philosophical Quarterly* 63(252):428–444.

Douven, Igor. 2021. Abduction. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Summer 2021 ed. Metaphysics Research Lab, Stanford University.

Dowding, Keith. 2015. *The philosophy and methods of political science*. Macmillan International Higher Education.

Dowding, Keith and Charles Miller. 2019. "On prediction in political science." *European Journal of Political Research* 58(3):1001–1018.

Dreber, Anna, David G Rand, Drew Fudenberg and Martin A Nowak. 2008. "Winners don't punish." *Nature* 452(7185):348–351.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts and Brandon M Stewart. 2022. "How to make causal inferences using texts." *Science advances* 8(42):eabg2652.

Fairfield, Tasha and Andrew E Charman. 2022. *Social inquiry and Bayesian inference: Rethinking qualitative research*. Cambridge University Press.

Fearon, James D and David D Laitin. 2003. "Ethnicity, insurgency, and civil war." *American political science review* 97(1):75–90.

Ford, Heather and Judy Wajcman. 2017. "'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap." *Social studies of science* 47(4):511–527.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Gelman, Andrew and Eric Loken. 2014. "The statistical crisis in science: data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up." *American scientist* 102(6):460–466.

Gelman, Andrew and Guido Imbens. 2013. Why ask why? Forward causal inference and reverse causal questions. Technical report National Bureau of Economic Research.

Gelman, Andrew and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gerber, Alan S, Donald P Green, Edward H Kaplan, Ian Shapiro, Rogers M Smith and Tarek Massoud. 2014. "The illusion of learning from observational research." *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* pp. 9–32.

Gerring, John. 2012. "Mere description." *British Journal of Political Science* 42(4):721–746.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Grimmer, Justin, Margaret E Roberts and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Gross, Justin H. 2015. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." *American Journal of Political Science* 59(3):775–788.

Harman, Gilbert H. 1965. "The inference to the best explanation." *The philosophical review* 74(1):88–95.

Heckman, James J and Burton Singer. 2017. "Abducting economics." *American Economic Review* 107(5):298–302.

Hempel, Carl. 1966. *Philosophy of Natural Science*. Englewood Cliffs, N.J: Prentice Hall.

Henderson, Leah. 2020. "Bayesianism and inference to the best explanation." *The British Journal for the Philosophy of Science* .

Huber, John. 2013. "Is Theory Getting Lost in the 'Identification Revolution'?" *The Money Cage* .

Huber, John D. 2017. *Exclusion by elections: inequality, ethnic identity, and democracy*. Cambridge University Press.

Humphreys, Macartan and Alan M Jacobs. 2015. "Mixing methods: A Bayesian approach." *American Political Science Review* 109(4):653–673.

Imai, Kosuke and Teppei Yamamoto. 2010. "Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis." *American Journal of Political Science* 54(2):543–560.

Ioannidis, John PA. 2005. "Why most published research findings are false." *PLoS medicine* 2(8):e124.

Keele, Luke, Randolph T Stevenson and Felix Elwert. 2020. "The causal interpretation of estimated associations in regression models." *Political Science Research and Methods* 8(1):1–13.

Kerr, Norbert L. 1998. "HARKing: Hypothesizing after the results are known." *Personality and social psychology review* 2(3):196–217.

King, Gary, Robert O Keohane and Sidney Verba. 1994. *Designing social inquiry*. Princeton university press.

Lau, Richard R. 1985. "Two explanations for negativity effects in political behavior." *American journal of political science* pp. 119–138.

Lee, David S. 2008. "Randomized experiments from non-random selection in US House elections." *Journal of Econometrics* 142(2):675–697.

Lieberson, Stanley and Joel Horwich. 2008. "Implication analysis: a pragmatic proposal for linking theory and data in the social sciences." *Sociological Methodology* 38(1):1–50.

Lipset, Seymour Martin. 1959. "Some social requisites of democracy: Economic development and political legitimacy1." *American political science review* 53(1):69–105.

Lipton, Peter. 2003. *Inference to the best explanation*. Routledge.

Little, Andrew T and Thomas B Pepinsky. 2021. "Learning from biased research designs." *The Journal of Politics* 83(2):602–616.

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86(3):532–565.

Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. University of Chicago Press.

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert and Jennifer L Tackett. 2019. "Abandon statistical significance." *The American Statistician* 73(sup1):235–245.

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Moore, Barrington et al. 1993. *Social origins of dictatorship and democracy: Lord and peasant in the making of the modern world*. Vol. 268 Beacon Press.

Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press.

Munger, Kevin, Andrew M Guess and Eszter Hargittai. 2021. "Quantitative description of digital media: a modest proposal to disrupt academic publishing." *Journal of Quantitative Description* 1(1):1–13.

Peirce, Charles. 1878. "How to make our ideas clear." *Popular Science Monthly* 12(Jan):286–302.

Psillos, Stathis. 2002. Simply the best: A case for abduction. In *Computational logic: Logic programming and beyond*. Springer pp. 605–625.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American journal of political science* 58(4):1064–1082.

Rosenthal, Robert. 1979. "The file drawer problem and tolerance for null results." *Psychological bulletin* 86(3):638.

Salmon, Wesley C. 1971. *Statistical explanation and statistical relevance*. London: University of Pittsburgh Press.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

Schnall, Simone, Jonathan Haidt, Gerald L Clore and Alexander H Jordan. 2008. "Disgust as embodied moral judgment." *Personality and social psychology bulletin* 34(8):1096–1109.

Sekhon, Jasjeet S. 2009. "Opiates for the matches: Matching methods for causal inference." *Annual Review of Political Science* 12:487–508.

Sekhon, Jasjeet S and Rocio Titiunik. 2012. "When natural experiments are neither natural nor experiments." *American Political Science Review* 106(1):35–57.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22(11):1359–1366.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman and Wolf Vanpaemel. 2016. "Increasing transparency through a multiverse analysis." *Perspectives on Psychological Science* 11(5):702–712.

Tavory, Iddo and Stefan Timmermans. 2014. *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Thompson, William R. 1975. "Regime vulnerability and the military coup." *Comparative Politics* 7(4):459–487.

Tukey, John W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford University Press.

Woodward, James. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.