Temporal Validity as Meta-Science

Kevin Munger*

March 20, 2023

Abstract

The "credibility revolution" has forced quantitative social scientists to confront the limits of our methods for creating general knowledge. As a result, many practitioners aim to generate valid but local knowledge and then synthesize and apply that knowledge to predict what will happen in a target context. Positivist social science has until recently been hamstrung with other, more immediate threats to validity and inference, but I argue that recent advances in statistical approaches to the problem of external validity reveal limits of the current paradigm. This article and the term "temporal validity" illustrate the intrinsic limits of agnostic (that is, assumption-free) external validity when the target setting is in the future. These limits, I argue, suggest a re-orientation of social science methodology. We should acknowledge that no research design, no empirical knowledge, is *perfectible*; instead, we should explicitly aim to increase the *amount* and *quality* of knowledge we produce. I believe it is useful to characterize this perspective as "Meta-Science," an emerging social/intellectual movement within the social sciences. "Temporal validity" and the implied "knowledge decay" thus represent a meta-scientific intervention aimed at increasing the usefulness of the knowledge we produce. Among other structural reforms, I argue that the binary reality of academic scholarship (a paper is published or not) reifies the perfectibility of empirical knowledge and is thus an impediment to recognizing the continuous nature of all forms of scientific validity.

*kevinmunger@gmail.com

1 The Imperfectibility of Social Science

> "We are directed in time, and our relation to the future is different from our relation to the past. All our questions are conditioned by this asymmetry, and all our answers to these questions are equally conditioned by it." – Norbert Wiener, *Cybernetics*.

Methodological progress has demonstrated that conducting internally valid empirical research—making true statements even about particular cases—is more difficult than once thought (Angrist and Pischke, 2010). The rise of randomized control trials (RCTs), regression-discontinuity designs and natural experimental approaches has increased the credibility of social science research, but it has also increased the relevance of concerns about external validity. Compared to regressions that purport to describe global phenomena, this research generates an internally valid estimate of the causal effect of a given treatment, in a given time and place, and on a given subject population (Samii, 2016).

The goal of this research is to accumulate generalizable knowledge; in the words of Dehejia, Pop-Eleches and Samii (2021), "with a large number of internally valid studies across a variety of contexts, it is reasonable to hope that researchers are accumulating generalizable knowledge...The success of an empirical research program can be judged by the diversity of settings in which a treatment effect can be reliably predicted" (p217).

This view is prominent across the social sciences. Political scientist Phil Schrodt argues that "explanation in the absence of prediction is not scientifically superior to predictive analysis, it isn't scientific at all!" (Schrodt, 2014). In Economics, 2019 Nobel Laureate Esther Duflo called for economists to "adopt the mindset of a plumber. Plumbers try to predict as well as possible what may work in the real world" (Duflo, 2017). Sociologist Duncan Watts also advocates for a more "solution-oriented social science" (Watts, 2017) and argues that prediction is a valuable tool for explanation (Hofman et al., 2021). And in Psychology, Tal Yarkoni has advocated for incorporating more prediction Yarkoni and Westfall (2017), which has in turn caused him to confront what he calls the "Generalizability Crisis" Yarkoni (2022).

My premise is thus that one of the primary goals of social science is *prediction*. There are other goals, of course, and social science is no stranger to methodological pluralism. See Ashworth, Berry and de Mesquita (2021) for a recent reflection on the state of the deductive paradigm, Mahoney (2021) for a discussion of the set-theoretic, qualitative paradigm, and Reiss (2007) for a general defense of pluralism in social science goals.

The "credibility revolution" taking place across the social sciences means that more attention must be paid to causal identification, and thus that researchers must devote more effort developing their identification strategy and less on novel theorizing. In a landmark outline of the paradigm he calls "causal empiricism," Samii (2016) argues that this does not mean that "'theory is being lost' but rather that theory is being held constant as we go about the difficult business of trying to do credible causal inference," and that "generalization and theory development are better left to synthesis studies."

The rate of production of internally valid causal knowledge is increasing, even accelerating, in the presence of methodological innovations, exogenous information technological developments and economies of scale in research practices. This is a major step forward, but it highlights a blind spot in the way that social science methods have been adapted towards this goal. Academic research takes place as time advances. As internally valid studies accumulate, the world changes. If this approach to social science is to succeed, it is necessary that the rate of knowledge accumulation outpace the rate at which old knowledge decays due to the world changing.

Knowledge "decays" if was produced in a temporal context that is relevantly distinct from when it is to be applied. That is, actions based on the knowledge would have originally aided the actor in achieving their desired ends; as the knowledge decays, it becomes less helpful, irrelevant or possibly even harmful.

The rate of knowledge decay is thus related to the *rate of change* of the world, what I will call r. I expand on this concept below, but consider r the change in the accuracy a machine learning algorithm gains from having access to that knowledge in making predictions in different temporal contexts. That is, if the predictor gains 2% accuracy from the knowledge in time T but gains only 1% accuracy in time $T + 1$, then r is 1.

This paper conceptualizes the problem of knowledge decay as a specific form of external validity: *temporal validity*. Although time-related scientific validity concerns date back at least to Hume, and although fields like time series econometrics and corpus linguistics have long centered change over time, I argue that the contemporary external validity literature that aims to accumulate, integrate and apply causal social scientific knowledge has not yet come to terms with the fact that the paradigm under which they operate implies application *in the future*. Prediction that does not acknowledge the scientist's temporal position—all of our knowledge comes from the past, all of the contexts to which we wish to apply that knowledge are in the future—is merely

counterfactual history.[1]

I use the term "temporal validity" not to claim any novel insight in either the philosophy of science or statistical methods, but simply to call attention to the inherent contradiction between existing methods for external validity, the inherited institutions of social science practice, and a paradigm aiming to make predictions.

All of the knowledge contained in published papers comes from the past, and all action is in the future: without confronting this reality, papers that perfectly explain the past are an exercise in what machine learning practitioners call "overfitting." Despite the goal of reducing the prediction error of estimates of the effects of human behavior in the "test dataset" of the future, current methodological efforts focus too heavily on minimizing bias (that is, maximizing internal validity) in the "training dataset" of the past.

Although not an explicit part of the research programme, I argue that the motivating spirit of the causal revolution—the agnostic perfectibility of internal validity through research design—shapes the orientation to the problem of external validity as well. As philosopher of social science Julian Reiss puts it, "External validity is normally juxtaposed with internal validity, and the former defined in terms of the latter." (Reiss, 2019)

More concretely: Egami and Hartman (2022) provide a rigorous framework for the conditions under which external validity is feasible. The structure of this argument extends the formalism of the Neyman–Rubin potential outcomes model at the heart of causal empiricism, consolidated in textbooks like Angrist and Pischke (2009).[2] The bulk of this impressive effort is devoted to defining the necessary assumptions and developing statistical machinery that would allow an internally valid inference to also be externally valid.

Making the necessary assumptions explicit reveals that they are generally implausible, at least when the researcher wishes to generalize results with the level of precision

[1]Other paradigms within social science have embraced the challenges posed by time. Pierson (2011) argues that path dependency and various forms of feedback make inference about the long-term effects of institutions especially challenging, and Grzymala-Busse (2011) makes related points about the importance of sequence and other issues of temporality in questions of which causal mechanisms come into play in periods of institutional transition.

[2]An alternative contemporary approach to causal reasoning, Directed Acyclic Graph (DAG) model developed by Judea Pearl has, in estimation, been less central to the development of causal empiricism. Although perhaps more useful for knowledge synthesis in a fully mature causal empiricism (Bareinboim and Pearl, 2016), the formalization of DAGs does not offer a solution to the issues I discuss in this paper. Indeed, I suspect that the tidy visualization of a DAG is more likely to inspire false certainty about causal sufficiency when applied to social science (Halpern, 2015).

currently valorized in normal scientific practice. Egami and Hartman (2022) pragmatically suggest a radically lower bar for external validity, abandoning the standard "effect-generalization" for "sign-generalization":

"Sign-generalization is also sometimes a practical compromise when effect-generalization is not feasible...when required data on target populations, treatments, outcomes, or contexts are not available."

Given that the required data on contexts can never be fully available when the target is in the future, effect-generalization is never entirely "feasible."

This conception of what I call "agnostic" external validity as a formal statistical procedure is the logical extension of the central emphasis of the causal empiricist paradigm: causal identification. RCTs are considered the gold standard research design because they are fully agnostic: they ensure unconfoundedness and positivity (internal validity) through research design, with zero modeling assumptions about the structure of the world (Aronow et al., 2021).

If r is positive, if the relationship between causally relevant variables changes over time — or if new values of those variables appear — then agnostic external validity is impossible if the target context is in the future. The problem of "temporal validity" is an insurmountable event horizon for agnostic social scientific knowledge. The unknowability of the future makes external validity imperfectible.[3]

The severity of the problem in practice varies, with r, across different realms of inquiry; for most social science questions, there are more pressing sources of error. However, decades of methodological progress have reduced many of these error sources. It is only by standing on the shoulders of giants that the problem of temporal validity becomes more than a third-order concern.

If external validity is imperfectible and all we can aim to achieve is sign-generalization, social science methodology and practice should be radically restructured. The question of how to restructure it is a meta-scientific one, with elements both normative—what, exactly, are we trying to achieve?—and positive—how do we evaluate our achievement?

Although the motivation for the causal empiricist paradigm is pragmatic, I agree with Hofman, Sharma and Watts (2017) that this pragmatism requires more rigor in its application: we need to make many more predictions (*about the future*) and standardize

[3]There are other notable approaches to external validity that do not operate within the agnostic paradigm and indeed which do not aim to make predictions at all; see Findley, Kikuta and Denly (2021) for a recent review. These approaches already invoke the necessity of "theoretically-guided...principled extrapolation" and are thus more aligned with my suggestions; still, I believe that this tradition would benefit from more explicit attention to the rate of change r of the various phenomena under study.

the process by which these predictions are evaluated. Temporal validity requires taking time seriously, and this includes using the veil of the future as both the strictest and most realistic test of the success of our knowledge production endeavors.

We should not abandon the goal of improving political decision-making by increasing the diversity of settings in which we can predict outcomes. At present, however, too much of our methodological innovation and too great a percentage of researcher energy and graduate curriculum is spent reinforcing the internal validity of individual research projects.

To evaluate where this energy should be re-allocated requires the embrace of meta-science as a major branch of social science methodology. We are not trying to produce perfect academic papers; producing perfect academic papers is impossible. Instead, we are trying to produce *knowledge*, and my normative stance is that we should be trying to produce knowledge that helps align human action with human intention—that is, knowledge that enhances the likelihood that the actions that people choose to take will result in their desired outcome, a *prediction*.

From any paradigmatic position, I argue that the fundamentally binary nature of academic publication (published or unpublished valid or not valid) and in-line citation (one unit of knowledge per citation, accumulating and not decaying) serve to reify the perfectibility of social science.

2 External Validity in Theory

> "No man ever steps in the same river twice, for it's not the same river and he's not the same man." – Heraclitus.

Given that a single study is necessarily restricted in its scope, the premise of the causal empiricist paradigm is to aggregate findings from across many studies such that the union of their scope covers the entirety of the covariate and treatment component spaces. This knowledge aggregation is not trivial, however, and this subject has been the focus of much recent methodological attention (Athey and Imbens, 2016; Dehejia, Pop-Eleches and Samii, 2021; Egami and Hartman, 2018; Gechter, 2015; Green and Kern, 2012; Hartman et al., 2015; Ho et al., 2007; Imai, Ratkovic et al., 2013; Kern et al., 2016; Nguyen et al., 2017; Phan et al., 2021; Slough and Tyson, 2022*a*; Stuart et al., 2011; Taddy et al., 2016; Wager and Athey, 2017).

However, the majority of the developments cited above use agnostic methods. One approach is to use matching and reweighting to find the location where the treatment effect is known that is most similar to the target context in the treatment $*$ covariate space. A complementary approach uses machine learning to discover the covariate values with the largest effect heterogeneity, restricting the space in which matching or reweighting is necessary; this is related to the "Double Machine Learning" approach of Chernozhukov et al. (2017), recently advanced within political science by Ratkovic (2022).

These statistical innovations reduce the costs of precise generalizability by identifying the portion of the covariate $*$ treatment component space for which we need internally valid estimates of treatment effects in order to transport those effects to the entire space.

The fundamental problem of generalizability to the future, then, is not that the covariate $*$ treatment component space is large, but that it *expands over time*. Because all of our knowledge is from the past and all the contexts to which we hope to apply that knowledge are in the future, it is always possible that Hotz, Imbens and Mortimer (2005)'s "no macro-effects assumption"/"support condition" will fail to obtain.

Let t denote the time at which a study is conducted, where $t < 0$ denotes the past, $t = 0$ the present, and $t > 0$ the future. Let X denote the covariate space of the units of analysis. Because time is unidirectional,

$$X_{t<0} \subseteq X_{t>0}$$

Time is infinitely divisible, and this process is not instantaneous. Let the *rate of change* (r) of a given phenomenon be the minimum time difference such that the covariate set expands:[4]

$$X_t \cup X_{t+r} \neq X_t$$

For a concrete example, consider the case of the Arab Spring. Ignoring the simplistic argument that Twitter was a necessary and sufficient cause of the revolutions

[4]Further assume that the covariate set expands in such a way that the value of the novel covariate cannot be predicted by other covariates:

$$x_{ij} \perp x_{i0}, x_{i1}, ..., x_{iC}$$

for some $x_{ij} \in X_t \cap X_{t+r}$

in the Middle East in the early 2010s,[5] it is undeniably the case that the spread of information technology changed the nature of the strategic problem facing regimes and protesters (Tufekci, 2017). In particular, government censorship used to cause a favorable media ecosystem. Once enough people had smartphones and internet access, this causal relationship failed to hold; the effect of government censorship on the success of a revolution became unclear.

In all of the "studies" conducted prior to the mid-2000s, the value of the covariate *Smartphone* was undefined.[6] After the exogenous spread of smartphones, however, *Smartphone* takes some non-zero value, violating the support condition. The rate of change r of the effect of government censorship was thus, around the time of the revolution in mobile phone technology, *high*.

2.1 Alternative Formalization

An agnostic formulation of the problem might be somewhat different. What are the necessary assumptions of a research design which actually *would* enable agnostic externally valid inference?

Define time T as a member of the covariate space X. One crucial element of the potential outcomes framework is the exclusion restriction:

$$Y \perp D|X$$

Where D is the treatment condition. Assume that we are discussing a perfectly designed and executed RCT, so that the above assumption holds; this is internal validity. Egami and Hartman (2022) extends this to external validity (in this simplified framework that discusses only Context (C)-Validity) and thus the contextual exclusion restriction:

$$Y_i(D = d, c) == Y_i(D = d, c^*)$$

Where c is the *context* in which the original study was run and c^* is the target context. How do we know if this is plausible? In a trivial sense, every context is

[5]There were many other social scientifically relevant mechanisms at play during the Arab Spring; for example, the bounded rationality noted by Weyland (2012) and the preference falsification described by Kuran (1991).

[6]Alternatively, if we define the covariate space as infinitely large, we can say that there was no variation in *Smartphone* in this time period, as it always took the value 0.

different, but how can we decide if a difference affects our potential outcomes?

Egami and Hartman (2022) define M_i as the vector of potential context moderators, allowing them to define the contextual exclusion restriction as:

$$Y_i(D = d, M = m, c) == Y_i(D = d, M = m, c^*)$$

This means that the potential outcomes conditional on treatment and context are the same, provided that the context moderators are the same across contexts. "This assumption is plausible when the measured context-moderators capture all the reasons why causal effects vary across contexts" (p13).

Reformulating this assumption to emphasize the time component:

$$Y_i(D = d, M = m, c_t) == Y_i(D = d, M = m, c_{t+r})$$

This requires that we have measured at time t (the present) all of the moderators of potential outcomes at time $t + r$ (the future). In other words: nothing matters in the future that does not matter today.

Returning to the case of the Arab Spring: what would an externally valid estimate of the effect of regime repression on the success of a revolution require, if the study was conducted in 2000 and the target context was the year 2011? It would require that the context-moderator $Smartphone_{2011}$ have been measured in the year 2000.

This is impossible.

2.2 The Necessity of Qualitative Knowledge

The use of mathematical notation in the previous section is begging the question at issue. The agnostic approach has *nothing* to say about the process by which new variables are added to the dataset. The world is always changing, and we must decide how to measure novel phenomena. How does it happen that a social scientist elects to add the *Smartphone* column to the covariate matrix? The current paradigm pays very little attention to this question, and full attention to violations of the exclusion restriction in a given study.[7]

The already-existing columns in that covariate matrix cannot answer this question. The necessary information is instead encoded in the variable *names*, which themselves

[7]Other paradigms in social science embrace this question as foundational; see Mahoney (2021) on the set-theoretic approach.

represent prior work by a human mind to decide that the numbers in that column refer to a relevant social phenomenon. There is no agnostic statistical process for determining the relevance of future social phenomena based on the "training data" from the past; this is a qualitative task.

Intuition and common sense imply that *some* of our knowledge must be transferable from existing covariates to the novel phenomenon—we might manipulate the covariates that are "most similar" to the novel covariate. This appeal to "similarity" or "relevance" is ultimately unavoidable. The philosopher of science Nancy Cartwright has repeatedly criticized RCTs on the grounds that generalizability ultimately requires some appeal to the target context being "similar enough" to known contexts (Cartwright, 2007*a*,*b*; Deaton and Cartwright, 2018).[8]

My critique is related: *agnostic approaches cannot achieve temporal validity.* Because time is unidirectional, the future will contain novel states of the world or novel treatment moderators that a given model *cannot* account for, regardless of how much data from the past it has access to. This was a common critique among philosophers of science, natural scientists, and qualitative or historical social scientists when the modern "naturalist" social sciences emerged in the postwar United States.[9] The goal of finding some "Unity of the Sciences" was doomed at the start: "[Behavioral scientists] consider that the main task of the immediate future is to extend to the fields of anthropology, of sociology, of economics, the methods of the natural sciences, in the hope of achieving a like measure of success in the social fields. From believing this necessary, they come to believe it possible. In this, I maintain, they show an excessive optimism, and a misunderstanding of the nature of all scientific achievement" (p162, Wiener (1948)).

Even the standard practice of social science is imperiled. Replication is generally considered a key component of this practice. But true replication—of all but the most tightly controlled lab or survey experiments—is impossible without some recourse to the a non-rigorous "similarity" between contexts. This is a serious problem for "naturalist" paradigm of social science.

Indeed, this problem has become sufficiently self-evident that social scientific reformers in other disciplines have resorted to *redefining the word "replication"* in order

[8]But see, among others, Imbens (2018), who argues that Cartwright's understanding is mistaken, or at a minimum that she and the applied statisticians she criticizes are talking past each other.

[9]Per the classification in Rosenberg (2018), "naturalist" as in "committed to methods adapted from the natural sciences" (p19).

to square the reality of social science with the naturalist tradition. Nosek and Errington (2020) argue that the "common understanding...of replication is intuitive, easy to apply, and incorrect." Instead, they assert that "Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research."[10]

This rhetorical move may lead to better scientific practice by sidestepping sometimes tedious debates about the whether one experiment is "similar enough" to another to count as a replication, but it requires giving up on the the agnostic approach to external validity entirely. This conception of replication represents a radical re-routing of the social scientific process through the intuitions and judgements of social scientists. Indeed, I argue that this brazen re-definition is evidence of a paradigm in distress (Kuhn, 2012).

3 External Validity in Practice

> "There can be no demonstrative arguments to prove, that those instances, of which we have had no experience, resemble those, of which we have had experience." – David Hume, *A Treatise of Human Nature*.

Frequently replicated experiments on a given population are insufficient, even in the presence of large sample sizes and rich individual-level covariate information. Allcott (2015) demonstrates this limitation in a paper on "site selection bias": even with "large samples totaling 508,000 households, 10 replications spread throughout the country, and a useful set of individual-level covariates to adjust for differences between sample and target populations." However, the "extrapolation bias" of the effect of the same intervention applied at other sites is an order of magnitude larger than the estimated standard error of the treatment effect. Similarly, Vivalt (2020) aggregates the results of impact evaluations of international development programs from 635 published papers. Development economics is "one of the first fields...with enough papers on comparable topics to do this analysis," and the results are not promising: "results are much more heterogeneous than in other fields."[11]

[10] Slough and Tyson (2022*b*) advance an alternative re-definition of "replication," one more amenable to social scientists working in the deductivist paradigm. I am similarly underwhelmed and argue that we should simply abandon the term "replication" to mark a more decisive split with the naturalist tradition.

[11] In a comparable paper from social psychology, Paluck, Green and Green (2018) perform a meta-analysis of the literature on the theory that inter-group contact reduces prejudice (Allport, 1954).

Tolerably unbiased extrapolation has been shown to be empirically possible. Frequently replicated experiments that span both decades and the globe can be used to aggregate treatment effects and extrapolate them to novel contexts. Using the Angrist and Evans (1996) natural experiment (that the sex distribution of a household's first two children acts as an as-if random assignment to have additional children), Dehejia, Pop-Eleches and Samii (2021) use 142 country-years of census data (with an aggregate sample size of 10 million) from the Integrated Public Use Microdata Series. In this case, knowledge appears to be accumulating: the addition of more country-years of data generally reduces the out-of-sample prediction error, and the use of a rich set of both micro- and country-level covariates reduces the out-of-sample prediction error to close to zero.[12] Bisbee et al. (2017) extends this approach to the case of instrumental variables.

Both of these cases require knowledge of the covariate values in the context being extrapolated to. Temporal validity reminds us that this is impossible, in practice. Rather than using the observed values of the relevant covariates (only possible because the "prediction" in these papers all takes place in the past), covariate-adjusted treatment effect prediction must first predict the value of the covariates. This additional source of variance is a problem, but not an insurmoutnable one; in the Dehejia, Pop-Eleches and Samii (2021) analysis, the most important covariates are macro-level variables like GDP per capita and total fertility rate, things that can be extrapolated at least a few years into the future with tolerable accuracy.

The more serious issue is that this approach cannot account for the creation of novel covariates or novel treatment moderators. Consider the recent COVID pandemic. This "pandemic" moderator does not exist in the dataset—that is, it was not measured as part of any of the natural experiments.

This discipline has not fully embraced field experiments, so they are only able to aggregate across 27 randomized field studies. The results are very different from the previous gold standard meta-analysis on the topic: Pettigrew and Tropp (2006) aggregates more than 500 studies and finds strong, context-independent and homogeneous effects of contact reducing prejudice. Restricted to the 27 well-conducted studies, however, Paluck, Green and Green (2018) find that these effects are in fact weaker, context-dependent and more heterogeneous. Even more troublingly, "not one study [of the over 500] assesses the effects of interracial contact on people older than 25." The lack of population sampling leaves open the possibility of far greater heterogeneity; although the results are not conclusive, the effect sizes of the studies conducted on adults over 25 were in general smaller than those on younger people.

[12]The authors admit that they cannot account for site selection into their database; all of the country-years share the property of "have data archived at IPUMS," and it is possible that the model would not extrapolate correctly to country-years which do not have this property.

But preliminary evidence suggests that the pandemic shifted the causal system under study. A report from the UN Population Fund finds that, beginning several months after March 2020, developed and middle-income countries experienced a significant decline in birthrates (UNFPA, 2021). Perhaps the effect of this macro-shock on the treatment effect of interest would be captured by the huge change in the covariate for total fertility—in which case the error would be introduced when predicting that covariate, which the veil of the future prevents us from observing directly.

It is possible that the problem is still more serious. Frueh (2022) argues for "a shift in who gave birth" in the United States, that the decline in fertility was disproportionately concentrated among older age groups and those with greater education and resources. This would imply that the relationship between these covariates and the treatment effect is different when $pandemic = 1$. This novel mechanism is not a source of error that covariate adjustment can fix.

A final problem is that the pandemic affected the literal data generating process—not the social phenomenon of interest, but rather the relationship between that phenomenon and the data produced by Census agencies. The magnitude of the shock of the pandemic on Census practices was high; the downstream implications on our capacity to predict the *the numbers they report* are unknown, and unknowable within the virtual, statistical world of agnostic external validity.

The authors argue that the external validity natural experiment under study is a "possibility result" (p236). Fair enough. But the logic of "site selection bias" suggests that we should not expect external validity results—the ones we observe, because they were conducted in a plausible context— to be externally valid with respect to the target population of relevant external validity results.

Even in this best-case scenario, I argue that r is both positive and unpredictable.

4 Conclusion

Unfortunately, this manuscript does not propose a solution to the problem of induction. That would be the ideal way out of what I see as a contradiction:

1. External validity requires knowledge of the target
2. The target context for prediction is in the future
3. We cannot have knowledge of the future

In other words: the agnostic impulse of causal empiricism—which I see as a dialectical response to the previous quantitative social science paradigm that embraced global but spurious causal claims based on researcher credulity and indeed hubris—remains, even as the methodological frontier moves beyond "local" internal validity. The contradiction emerges because agnostic external validity requires knowledge (*gnosis*).

Given the imperfectibility of our knowledge, the members of the community of social scientists whose primary goal is to increase the accuracy of predictive accuracy are forced to confront what I see as the fundamental question of meta-science: *how should we allocate our scare social science resources?*

Just as with standard social science, this meta-scientific question should be approached with a combination of quantitative and qualitative methods and then synthesized in institutions of knowledge production.

For example: the concept of temporal validity might (when further elaborated) help align social scientific resources and methods with the scope of possibility. I have argued that the amount of bias in future predictions is related to the rate of change r. At any given time, r varies across different subject areas.

As meta-scientific inquiry into temporal validity is not yet mature, my priors about where r is likely to be high or low are extremely diffuse (other than for online behavior, my area of substantive expertise). My goal in this manuscript is to convince other social scientists to begin taking r seriously, to begin developing qualitative tools to sharpen our intuitions about how r might vary across context and research question and developing quantitative tools for measuring r and then refining those intuitions.

For many of the subjects that have been studied with RCTs, r has been sufficiently low that its contribution to bias has been small relative to problems of experimental design and practical implementation, as well as the total variance of effect heterogeneity for a given treatment. That is, temporal validity is a less serious source of bias than, say, construct validity, in the context of development experiments.

However, more and more human behavior is taking place online, a context in which r is likely to be high. This makes certain methods a poor allocation of our scarce resources. Consider a hypothetical example. Assume that study XXXX (2014) and YYYY (2022) were both perfectly-executed RCTs in which a random sample of US Facebook users were paid to stop using Facebook for a month. We might then encounter the following sentences in a research article: "XXXX (2014) finds that Facebook desistance causes an increase in partisan affective polarization; however, YYYY (2022) finds the opposite, that Facebook desistance causes a decrease in partisan affective polarization. Future

research is needed to adjudicate which of these is correct."

The example is intended to be absurd: r is far too high this context, so that accumulating knowledge about the effect of "Facebook use" is meaningless. Another perspective on this problem is that "Facebook use" is a poorly-defined construct, that it bundles together too many disparate treatments and mechanisms. This is likely true, absent any efforts at construct validity (Esterling et al., 2021). But citizens and researchers alike think it is a meaningful construct, that it structures their understanding of the world. And the related policy questions are certainly salient: "Is Facebook good or bad? Should I use Facebook?"

This is not an argument for nihilism on the question of Facebook, however. Just as there are contexts in which effect-generalizability is implausible, I contend that contexts in which r is sufficiently high are *too expensive* for rigorous causal knowledge. Given our community's inelastic endowment of resources, I believe that pursuing this knowledge will not generate sufficiently durable knowledge to justify the cost. But there are many other scientific questions that social scientists can hope to answer: "What *is* Facebook? How has it changed over time? Who uses Facebook? How do they use it?"

The answers to qualitative or descriptive questions like these *can* be used to make predictions about causal effects in the future, even in the absence of causal estimates from the past—but only if we embrace the need for *humans in the loop*. I argue that agnostic temporal validity is a dead end. Instead of attempting to excise human subjectivity, we should take it more seriously. By "human subjectivity" here I mean both our formal priors as researchers and our human biases about, for example, what research questions are most important. More of the energy of social scientists—more of the attention of methodologists, more of our *rigor*—should be allocated towards studying our own beliefs and in using those beliefs to make predictions about the future.

A necessary first step is to *actually* make predictions about the future. There has been considerable interest in the methods for rigorous prediction over the past decade, using experts who invest intensively in the task (Tetlock, 2009), moderate-intensity surveys of practitioners, as in the Social Science Prediction Platform (DellaVigna, Pope and Vivalt, 2019), or more decentralized prediction markets (Arrow et al., 2008). The recent COVID prediction project reported in Golden et al. (2023) is especially promising.

These approaches provide some rigor at the end of the chain of knowledge production and application. But the middle of this chain—the normal science we have

inherited, with the goal of producing machine-readable knowledge that can be used for agnostic external validity—should be restructured as well. An anonymous reviewer asks "why Bayesian updating fails...why we can't gradually update our priors based on past knowledge when a new study comes out, which is implicitly what are doing now?"

My personal prior is that this Bayesian updating is performing poorly. As good Bayesians, we need to confront our priors with evidence. However, we have only begun to allocate much rigor to this crucial link in the knowledge production chain; see, for example, the pioneering work by Little and Pepinsky (2021). In other words: how much time do social scientists spend reading studies as they come out? Is this the right amount of time? How much heterogeneity is there in the priors of scholars in a given subfield? How do they update in response to new information? Absent rigorous, empirical evidence, I don't know how to answer these questions; this manuscript is a call for the development of this evidence.

References

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130(3):1117–1165.

Allport, Gordon Willard. 1954. *The Nature of Prejudice*. Basic Books.

Angrist, Joshua D and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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.

Angrist, Joshua D and William N Evans. 1996. Children and their parents' labor supply: Evidence from exogenous variation in family size. Technical report National bureau of economic research.

Aronow, PM, James M Robins, Theo Saarinen, Fredrik Sävje and Jasjeet Sekhon. 2021. "Nonparametric identification is not enough, but randomized controlled trials are." *arXiv preprint arXiv:2108.11342* .

Arrow, Kenneth J, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson et al. 2008. "The promise of prediction markets.".

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 Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Bareinboim, Elias and Judea Pearl. 2016. "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences* 113(27):7345–7352.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches and Cyrus Samii. 2017. "Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect." *Journal of Labor Economics* 35(S1):S99–S147.

Cartwright, Nancy. 2007*a*. "Are RCTs the gold standard?" *BioSocieties* 2(1):11–20.

Cartwright, Nancy. 2007*b*. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen and Whitney Newey. 2017. "Double/debiased/neyman machine learning of treatment effects." *American Economic Review* 107(5):261–65.

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

Dehejia, Rajeev, Cristian Pop-Eleches and Cyrus Samii. 2021. "From local to global: External validity in a fertility natural experiment." *Journal of Business & Economic Statistics* 39(1):217–243.

DellaVigna, Stefano, Devin Pope and Eva Vivalt. 2019. "Predict science to improve science." *Science* 366(6464):428–429.

Duflo, Esther. 2017. "Richard t. ely lecture: The economist as plumber." *American Economic Review* 107(5):1–26.

Egami, Naoki and Erin Hartman. 2018. Covariate Selection for Generalizing Experimental Results. Technical report Working Paper.

Egami, Naoki and Erin Hartman. 2022. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review)* .

Esterling, Kevin, David Brady, Eric Schwitzgebel et al. 2021. "The Necessity of Construct and External Validity for Generalized Causal Claims." *OSF Preprints* 27.

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2021. "External validity." *Annual Review of Political Science* 24:365–393.

Frueh, Sarah. 2022. *The Pandemic 'Baby Bust' and Rebound*. National Academies. **URL:** *https://www.nationalacademies.org/news/2022/06/the-pandemic-baby-bust-and-rebound*

Gechter, Michael. 2015. "Generalizing the results from social experiments: Theory and evidence from mexico and india." *manuscript, Pennsylvania State University* .

Golden, Miriam A., Alexandra Scacco, Haoyu Zhai, Tara Slough, Macartan Humphreys, Eva Vivalt, Alberto Dias-Cayeros, Kim Yi Dionne, Sampada KC and Eugenia Nazrullaeva. 2023. "Gathering, evaluating, and aggregating social scientific models of COVID-19 mortality." *Working paper* .

Green, Donald P and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly* 76(3):491–511.

Grzymala-Busse, Anna. 2011. "Time will tell? Temporality and the analysis of causal mechanisms and processes." *Comparative Political Studies* 44(9):1267–1297.

Halpern, Orit. 2015. *Beautiful data: A history of vision and reason since 1945*. Duke University Press.

Hartman, Erin, Richard Grieve, Roland Ramsahai and Jasjeet S Sekhon. 2015. "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3):757–778.

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15(3):199–236.

Hofman, Jake M, Amit Sharma and Duncan J Watts. 2017. "Prediction and explanation in social systems." *Science* 355(6324):486–488.

Hofman, Jake M, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire et al. 2021. "Integrating explanation and prediction in computational social science." *Nature* 595(7866):181–188.

Hotz, V Joseph, Guido W Imbens and Julie H Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125(1-2):241–270.

Imai, Kosuke, Marc Ratkovic et al. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Imbens, Guido. 2018. "Comments on understanding and misunderstanding randomized controlled trials: A commentary on Cartwright and Deaton." *Social science & medicine (1982)* .

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill and Donald P Green. 2016. "Assessing methods for generalizing experimental impact estimates to target populations." *Journal of research on educational effectiveness* 9(1):103–127.

Kuhn, Thomas S. 2012. *The structure of scientific revolutions*. University of Chicago press.

Kuran, Timur. 1991. "Now out of never: The element of surprise in the East European revolution of 1989." *World politics* 44(01):7–48.

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

Mahoney, James. 2021. *The logic of social science*. Princeton University Press.

Nguyen, Trang Quynh, Cyrus Ebnesajjad, Stephen R Cole, Elizabeth A Stuart et al. 2017. "Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects." *The Annals of Applied Statistics* 11(1):225–247.

Nosek, Brian A and Timothy M Errington. 2020. "What is replication?" *PLoS biology* 18(3):e3000691.

Paluck, Elizabeth Levy, Seth A Green and Donald P Green. 2018. "The contact hypothesis re-evaluated." *Behavioural Public Policy* pp. 1–30.

Pettigrew, Thomas F and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90(5):751.

Phan, My, David Arbour, Drew Dimmery and Anup Rao. 2021. Designing Transportable Experiments Under S-admissability. In *International Conference on Artificial Intelligence and Statistics*. PMLR pp. 2539–2547.

Pierson, Paul. 2011. Politics in time. In *Politics in Time*. Princeton University Press.

Ratkovic, Marc. 2022. "Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression." *American Political Science Review* pp. 1–17.

Reiss, Julian. 2007. "Do we need mechanisms in the social sciences?" *Philosophy of the social sciences* 37(2):163–184.

Reiss, Julian. 2019. "Against external validity." *Synthese* 196(8):3103–3121.

Rosenberg, Alexander. 2018. *Philosophy of social science*. Routledge.

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

Schrodt, Philip A. 2014. "Seven deadly sins of contemporary quantitative political analysis." *Journal of peace research* 51(2):287–300.

Slough, Tara and Scott A Tyson. 2022*a*. "External Validity and Meta-Analysis." *American Journal of Political Science* .

Slough, Tara and Scott A Tyson. 2022*b*. "Sign-Congruence, External Validity, and Replication.".

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw and Philip J Leaf. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386.

Taddy, Matt, Matt Gardner, Liyun Chen and David Draper. 2016. "A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation." *Journal of Business & Economic Statistics* 34(4):661–672.

Tetlock, Philip E. 2009. Expert political judgment. In *Expert Political Judgment*. Princeton University Press.

Tufekci, Zeynep. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

UNFPA. 2021. *How will the COVID-19 pandemic affect births?* Technical Brief.

Vivalt, Eva. 2020. "How much can we generalize from impact evaluations?" *Journal of the European Economic Association* 18(6):3045–3089.

Wager, Stefan and Susan Athey. 2017. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* (just-accepted).

Watts, Duncan J. 2017. "Should social science be more solution-oriented?" *Nature Human Behaviour* 1(1):1–5.

Weyland, Kurt. 2012. "The Arab Spring: Why the surprising similarities with the revolutionary wave of 1848?" *Perspectives on Politics* 10(4):917–934.

Wiener, Norbert. 1948. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.

Yarkoni, Tal. 2022. "The generalizability crisis." *Behavioral and Brain Sciences* 45.

Yarkoni, Tal and Jacob Westfall. 2017. "Choosing prediction over explanation in psychology: Lessons from machine learning." *Perspectives on Psychological Science* 12(6):1100–1122.