Research Note

Research and Politics
October-December 2024: 1–6

DOI: 10.1177/20531680241307942
journals.sagepub.com/home/rap
Sage

Gnostic notes on temporal validity

Austin Jang[1], Molly Offer-Westort[2], Serena Wang[3] and P. M. Aronow[4]

Abstract

Kevin Munger argues that, when an agnostic approach is applied to social scientific inquiry, the goal of prediction to new settings is generically impossible. We aim to situate Munger's critique in a broader scientific and philosophical literature and to point to ways in which *gnosis* can and, in some circumstances, must be used to facilitate the accumulation of knowledge. We question some of the premises of Munger's arguments, such as the definition of statistical agnosticism and the characterization of knowledge. We further emphasize the important role of microfoundations and particularism in the social sciences. We assert that Munger's conclusions may be overly pessimistic as they relate to practice in the field.

Keywords

Causal inference, agnosticism, external validity, temporal validity

Introduction

Munger (2023) provides a provocative discussion on the limits of *temporal validity* in the social sciences. In short, the article argues that, for an agnostic analyst unwilling to make strong assumptions, it is generally impossible to produce findings that are generalizable to the future.[1] Insofar as prediction is a central aim of the social sciences, Munger argues that a positivist approach will struggle to accumulate generalizable knowledge as it is inevitable that our knowledge base will decay and our ability to make useful predictions will suffer. Consequently, this argument implies that we should reorient toward a fundamentally meta-scientific strategy to knowledge production that emphasizes attention to variation in the rate of knowledge decay across subject areas and human subjectivity about the relative importance of scientific questions.

In our discussion, we aim to situate Munger's critique in a broader scientific and philosophical literature—including economics, anthropology, and the natural sciences—and point to ways in which *gnosis* can and, in some circumstances, practically must be used to facilitate the accumulation of knowledge.[2] We therefore raise questions about the premises of some of Munger's arguments, including the definition of agnosticism and characterization of prediction. We further emphasize the importance of microfoundations in the social sciences, both as an alternate means of generalizability and as a source of knowledge decay itself.

Agnosticism and causality

Munger describes agnostic inference as being "assumption-free." Specifically, randomized control trials (RCTs) are ascribed gold-standard research design status precisely

[1]Departments of Statistics and Data Science, Political Science, Yale University, USA
[2]Political Science, University of Chicago, USA
[3]Computer Science, Harvard University, USA
[4]Departments of Statistics and Data Science, Political Science, Biostatistics, and Economics, Yale University, USA

Corresponding author:
Austin Jang, Yale University, Departments of Statistics and Data Science, Political Science, 115 Prospect Street, New Haven 06520, USA.
Email: austin.jang@yale.edu

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" (Munger 2023: p. 3). Munger attributes this point of view to Aronow et al. (2021).[3] Although we agree that RCTs hold a special position among scholars due to the fact that they can facilitate inference under weaker assumptions than do other quantitative approaches, Munger's characterization of agnosticism as truly "assumption-free" is more akin to atheism. We instead conceptualize agnostic science as an aspirational approach to understanding our results under minimal assumptions, rather than one that avoids assumptions altogether.[4] While generalizability requires assumptions about unobservable model objects, so does the estimation of causal effects—even in ideal experimental conditions.

To contextualize this stance, we first note that standard analyses of RCTs do not make "zero-modeling assumptions."[5] Some assumptions are granted by the research design; for instance, an implication of uniform random assignment is that *in expectation*, unobserved covariates are balanced between the treatment and control group. However, to define—much less estimate—a causal effect, we still must make some assumptions about what information is revealed by treatment assignment. This is traditionally done through Rubin (1980)'s stable unit treatment value assumption (SUTVA), in which the treatment assigned to each unit is mapped to a single potential outcome. While often taken for granted, this is a strong assumption about "the structure of the world," as it implies no interference (i.e., one unit's treatment does not causally affect another's outcome) and no hidden treatment variations (i.e., a "consistency" assumption). Defining causal effects, like all structural parameters, requires an implicit or explicit modeling assumption that the parameters themselves are well defined. A definition of agnosticism that excludes assumptions excludes the study of causal effects.

Second, we frequently invoke assumptions in the natural world in settings where RCTs are infeasible if not altogether impossible. In fact, given the relative recency and limited adoption of RCTs, Rubin (1974) observed that most "scientific 'truths'" have been established without randomized experiments. Paul Holland's (1986) foundational paper on causal inference formalizes several sufficient assumptions for inferring causal effects, including the classical *statistical solution* of randomization. However, Holland also discusses *scientific solutions* that invoke alternate assumptions such as temporal consistency and unit homogeneity. Applied generally, these are strong assumptions about the way that the world works, but they are palatable and minimal with respect to their contexts: we assume we understand the causal effect of flicking off a light switch because the mechanism behind the switch itself is understood to be invariant under usual circumstances under which it is used. These assumptions may seem so minimal we forget we are making them, but an evidence-based approach to knowledge always requires engaging with some assumptions.

Returning to Munger's argument, we interrogate the premise that *agnostic* temporal validity is not possible because it would require making strong assumptions. To do so, we advance an approach to agnosticism that is better characterized as assumption-skeptical rather than assumption-free, and note that inferring causal effects will always require some set of strong assumptions. Indeed, any form of inference will require assumptions, whether time is included as a dimension of inference or not. This is not a hopeless predicament for the agnostic researcher; the challenge of finding a palatable set of assumptions is an important part of the scientific process.

How knowledge travels

Munger proposes an approach to knowledge "that helps align human action with human intention" (Munger 2023: p. 3). To advance this goal, Munger begins by engaging with Hume's problem of induction, an important starting point in grappling with the fundamental difficulty of moving from observations to inferences about yet-unseen cases. Munger proceeds to focus on the problem of extrapolating causal estimates to future populations, which is how Munger characterizes prediction.[6] However, one key obstacle to *temporal* validity is that the necessary context for prediction is unknown from the positivist perspective of the present. This forms the basis of Munger's unsolved contradiction (2023, p. 6):

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

We agree that if the only criteria for a successful research program is the guaranteed transportability of a causal parameter, then this contradiction would be troubling. However, this relies on a narrow conceptualization of what knowledge is and how we value it.[7] We instead argue that our accumulated knowledge may still prove *useful* even when we are unable to make causal predictions.

Rigorous causal knowledge may contribute to constructing heuristic models, even when the causal parameter fails to travel in a literal sense. Herbert Simon proposes that the aim of science is to seek parsimonious models "in the midst of apparent complexity and disorder." In developing such models, we are willing to trade off predictive fit with parsimony because there is value to "laws of qualitative structure" that "provide high-level generalizations, and representations useful in organizing problem-solving search" (Simon 2001: pp. 54–55, 70). In the natural sciences, "The germ theory of disease simply amounts to the

advice that: 'If you encounter a disease, look for a microorganism as cause; there may often be one.'" Even though this model is often wrong ("there are many diseases that don't involve microorganisms") such heuristic knowledge is both effective at communicating our scientific understanding of the world and useful for guiding real-world decisions. In economics, laws of qualitative structure such as utility theory and bounded rationality are a starting point rather than a culmination of scientific inquiry, serving as "scientists' hunting license" (Simon 2001, 57).

Simon's pattern-seeking is a form of generalizability, "To be able to sum up a complex body of data in a relatively simple generalization (a pattern) is to explain much with little" (Simon, 2001 p. 33). We may also gain knowledge through investigating and attempting to understand particulars before attempting inductive generalization. Even when a causal parameter estimated in one setting fails to travel to another, we may be able to learn sufficiently about local dynamics so that models we fit in one setting will still be useful in other settings. As well, capturing estimates of quickly changing causal effects before they decay may be especially important so that we can learn from these contextual particulars later—indeed, quickly changing systems are often of great interest to social scientists.

Historical particularism, popularized by Boas (1920) in anthropology, emphasizes the study of the specific historical contexts and processes that shape societies. Studying unique dynamic processes may then shed light on commonalities in conditions: "the method which we try to develop is based on a study of the dynamic changes in society that may be observed at the present time. We refrain from the attempt to solve the fundamental problem of the general development of civilization until we have been able to unravel the processes that are going on under our eyes." A key principle of this approach is that the contextual and historical understanding of specific phenomena will contribute to our understanding of microfoundations, the dynamics which emerge from social, psychological, and physiological conditions.

Munger argues that "contexts in which [decay] is sufficiently high are *too expensive* for rigorous causal knowledge." We cannot agree. The researcher's objective function may place value on other results beyond stable causal effect estimates. Rigorous causal knowledge, even when quickly decaying, can play a role in building laws of qualitative structure as characterized by Simon, or, in the vein of Boas, establishing the details of particular local dynamics.

Munger also asks us to embrace the need for "humans in the loop" and to "take [human subjectivity] more seriously." We believe that part of doing so must entail (1) recognizing the distinct desires of individual scientists to pursue particular research questions, regardless of their risk of knowledge decay and (2) the humility that we, as a field, do not always have the foresight to know ex ante which results will be most instrumental ex post.[8] Our proposal is not to dismiss the importance of resource considerations in research strategy, but rather reflects a belief that scientific knowledge produced by researchers with diverse objectives will ultimately result in a more nuanced and richer understanding of the world around us. Indeed, Boas' version of anthropological agnosticism would seem to reject relying on strong priors about how the world works for the sake of strategizing scientific study.[9]

Why knowledge decays

In formalizing the idea of knowledge decay, Munger provides the concept of a decay rate *r*, which, even in a "best case scenario… is both positive and unpredictable" (p 6). Two examples of sources of decay discussed by Munger are the technological shock of the invention of social media and the unexpected COVID pandemic. We agree that sometimes the world changes in ways that are completely unforeseeable given current information. However, when studying social phenomena, many changes are better described as adaptations or learning. Such processes may not be completely mysterious to the researcher and the underlying microfoundations may themselves reasonably be objects of study.

Turning towards the literature on economic forecasting, Robert Lucas (1976) contended that models validated in the short-run could fail in long-run prediction of causal contrasts due to drifts in key parameters, denoted by parameter drift of *theta*. The theta parameter drift problem mirrors that of Munger's *r*, decay rate. Where Lucas's approach deviates from Munger's is in offering a constructive account for parameter drift predicated on microfoundations; in Lucas' view, a primary driver of parameter drift is agent-based learning. While the covariate distributions, and their effect on key outcomes, can change, these shifts are often caused by agent adaptations, which are not completely opaque to the researcher. Lucas asserts that, "[A]gents' responses become predictable to outside observers only when there can be some confidence that agents and observers share a common view of the nature of the shocks which must be forecast by both." This is far from a modest condition, but it does provide guidance for a path forward: the inspection of incentives of relevant actors and the strategies available to them. Even in settings where causal parameters are unstable, we can design causal studies to learn about microfoundations. Closer alignment between formal theory and experiments may help (see, e.g., Ashworth et al., 2021).

Similar constructive accounts for the nature and origins of parameter drift populate the social sciences. One prominent example is Goodhart, 1984: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."[10] A well-known manifestation of this has been shown in public health,

where publishing surgery mortality rates is believed to have led to worse outcomes for sicker patients in part because reported mortality rates rewarded hospitals for treating healthier patients (Dranove et al., 2003). Using Munger's own example of Facebook use and knowledge decay, such technology platforms are highly responsive to study and measurement: benchmarks for measuring the performance of artificial intelligence (AI) and machine learning systems such as the Massive Multitask Language Understanding benchmark (MMLU) (Hendrycks et al., 2021) were developed by academic researchers, but companies have a strong incentive to adapt their models specifically to perform well on these highly publicized benchmarks. This has created a new frontier where researchers study the effects of this adaptivity on the validity of benchmarks as a measure of model capability (Dominguez-Olmedo et al., 2024; Recht et al., 2019). Similarly, content creators on platforms continuously adjust their behavior to fit algorithmic feeds, leading to strategic content production that shifts based on the platform's ranking systems. As a result, studies on the effects of algorithmically ordered as compared to chronologically ordered feeds must account for the dynamic interaction between platform algorithms and creator behavior, where the content itself evolves in response to the treatment. In the machine learning literature, a sub-field has emerged that studies strategic adaptations to machine learning algorithms deployed in consequential decision-making contexts, such as loan allocation and hiring (see, e.g., Hardt et al., 2016; Perdomo et al., 2020; Björkegren et al., 2020; Hardt et al., 2023). One can imagine how Goodhart's Law can manifest in political phenomena such as aid conditionality, classification of economic development, and human rights violations. However, the fact that agents respond to the incentives created by a measure should not dissuade researchers from the pursuit of careful measurement—the incentives generated by measurement and consequent responses are themselves important areas of study.

A defining feature of the social sciences is that we study people, and people as individuals and communities strategize, change, and adapt. That a social process is dynamic, even changing quite rapidly, should not deter researchers from causal inquiry of such a process for fear their results will become too quickly obsolete. It is a reasonable position that these very dynamics ought to take center-stage, so we can understand the world around us even as it changes.

Concluding remarks

While we have chosen to explore Munger's conception of temporal validity primarily through the lens of the philosophy of science, there are also many technical, statistical innovations that advance the goal of studying changing systems. Among these are: extensions of traditional sensitivity analysis, designs for non-stationary multi-arm problems, and distributionally and adversarially robust learning. We do not, like Munger, find a contradiction among "existing methods for external validity, the inherited institutions of social science practice, and a paradigm aiming to make predictions"—indeed, we are optimistic that with a bit of *gnosis*, we can make useful and reasonable claims about the conditions under which our findings may generalize.

Acknowledgments

The authors thank Fredrik Sävje for thoughtful comments and feedback.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Austin Jang https://orcid.org/0009-0006-5344-1354

Notes

1. We note that, when formally defined, the terms generalizable and transportable are distinct, where generalizability refers to extending causal knowledge from the sample to the population it was drawn from, while transportability refers to extending causal knowledge to the target population that is at least partly external to the original population. See Degtiar and Rose (2023) or Findley et al. (2021) for a review. This distinction is not germane to the points we make and we use the language of generalizability when discussing temporal validity.
2. By gnosis we mean "spiritual" (i.e., at least in part non-verifiable) beliefs about the generating model and related dynamics.
3. In a private correspondence, Munger clarified that this text was designed to reflect the usual argument for the "gold standard" position of RCTs rather than being Munger's personal view, attributing this view to Aronow et al. (2021). We disagree with this characterization of Aronow et al. While the paper does assert a special position for RCTs, this argument is predicated on the fact that an unbiased and uniformly consistent estimator may exist in RCTs due to knowledge of the propensity score (Robins and Ritov 1997), whereas the same is not generally true for observational studies even when unconfoundedness and positivity are assumed to hold. Thus, Aronow et al. explicitly sets aside the identification concerns that Munger adduced for the usual gold-standard argument.

4. Following Lin (2013)'s "Agnostic notes on regression adjustments to experimental" literal belief in assumptions is not requisite to learning from settings in which they are applied: "One does not need to believe in the classical linear model to tolerate or even advocate OLS adjustment, just as one does not need to believe in the Four Noble Truths of Buddhism to entertain the hypothesis that mindfulness meditation has causal effects on mental health."
5. Although defining and estimating causal effects both require structural assumptions, it is remarkably possible to *test* against null hypotheses about claims in particular causal models without invoking any structural assumptions in RCTs. This framework, built around Fisher's Exact Test, forms the basis of *randomization inference* (see, e.g., Rosenbaum 2002: Chapter 2). But when we seek to affirmatively characterize any causal effects, we are reliant on precisely these causal models to make headway. Thus, while the space of plausible hypotheses can be narrowed by RCTs without structural assumptions, we are nevertheless limited in the utility of this result for the advancement of any positivist research agenda.
6. Munger's emphasis on temporal validity that applies "causal social scientific knowledge…*in the future*" implies a definition of prediction that is oriented towards a causal parameter of a model. In the typical use of the term in statistics and machine learning, *prediction* is defined directly with respect to an outcome, conditional on predictors (see, e.g., Hardt and Recht 2022; Hastie et al., 2009). However, a focus on the external validity of causal parameters themselves is not novel in the social sciences (Campbell, 1957).
7. Munger acknowledges that beyond prediction, "There are other goals, of course, and social science is no stranger to methodological pluralism (p 2)."
8. The history of basic science is replete with examples of results with no immediate practical relevance forming the basis of future innovation decades later. Einstein himself was emphatically skeptical of the eventual possibility of using the atom to generate energy Moszkowski (1921/2014).
9. We are not the first to consider Boas's approach "agnostic," see Stocking Jr. (1966).
10. A more colloquially known version of Goodhart's Law is "when a measure becomes a target, it ceases to be a good measure," which is attributable to Strathern (1997).

References

Aronow PM, Robins JM, Saarinen T, et al. (2021) Nonparametric identification is not enough, but randomized controlled trials are. arXiv preprint arXiv:2108.11342.

Ashworth Scott, Berry Christopher and Bueno de Mesquita Ethan (2021) *Theory and credibility: Integrating theoretical and empirical social science*. Princeton University Press.

Björkegren D, Blumenstock JE and Samsun K (2020) "Training machine learning to anticipate manipulation". arXiv preprint arXiv:2004.03865.

Boas F (1920) The methods of ethnology. *American Anthropologist* 22(4): 311–321.

Campbell DT (1957) Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* 54(4): 297–312. DOI: 10.1037/h0040950.

Degtiar I and Rose S (2023) "A review of generalizability and transportability". *Annual Review of Statistics and Its Application* 10(1): 501–524.

Dominguez-Olmedo R, Dorner FE and Hardt M (2024) Training on the test task confounds evaluation and emergence. arXiv preprint arXiv:2407.07890.

Dranove D, Kessler D, McClellan M, et al. (2003) Is more information better? the effects of "report cards" on health care providers. *Journal of Political Economy* 111(3): 555–588.

Findley MG, Kikuta K and Denly M (2021) External validity. *Annual Review of Political Science* 24(1): 365–393.

Goodhart CAE (1984) Problems of monetary management: the UK experience. In: *Monetary Theory and Practice*. Berlin: Springer, 91–121.

Hardt M and Recht B (2022) *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton: Princeton University Press.

Hardt M, Megiddo N, Papadimitriou C, et al. (2016) Strategic classification. In: Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016, 111–122.

Hardt M, Mazumdar E, Mendler-Dünner C, et al. (2023) Algorithmic collective action in machine learning. In: *International Conference on Machine Learning*. Westminster: PMLR, 12570–12586.

Hastie T, Tibshirani R, Friedman JH, et al. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, Vol. 2.

Hendrycks D, Burns C, Basart S, et al. (2021) Measuring massive Multitask Language Understanding. In: International Conference on Learning Representations, Singapore, Thu Apr 24 – Mon Apr 28th, 2025.

Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81(396): 945–960.

Lin W (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Annals of Applied Statistics* 7(1): 295–318.

Lucas RE Jr (1976) Econometric policy evaluation: a critique. *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.

Moszkowski A (2014). Einstein the searcher: his work explained from dialogues with einstein. In: *Translated from the Original German by Henry L. Brose*. Oxfordshire: Routledge.

Munger K (2023) Temporal validity as meta-science. *Research & Politics* 10(3): 20531680231187271.

Perdomo J, Zrnic T, Mendler-Dünner C, et al. (2020) Performative prediction. In: *International Conference on Machine Learning*. Westminster: PMLR, 7599–7609.

Recht B, Roelofs R, Schmidt L, et al. (2019) Do ImageNet classifiers generalize to ImageNet? *Proceedings of the*

36th International Conference on Machine Learning, in Proceedings of Machine Learning Research 97: 5389–5400, Available from: https://proceedings.mlr.press/v97/recht19a.html.

Robins JM and Ritov Y (1997) Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* 16(3): 285–319.

Rosenbaum PR (2002) *Observational Studies*. New York: Springer.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5): 688–701.

Rubin DB (1980) Randomization analysis of experimental data: the Fisher randomization test comment. *Journal of the American Statistical Association* 75(371): 591–593.

Simon HA (2001) Science seeks parsimony, not simplicity: searching for pattern in phenomena. In: *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. Cambridge: Cambridge University Press, 32–72.

Stocking GW JR (1966) Franz Boas and the culture concept in historical perspective. *American Anthropologist* 68(4): 867–882.

Strathern M (1997) 'Improving ratings': audit in the British University system. *European Review* 5(3): 305–321.