2023

Knowledge Generation and Uncertainty in an Unpredictable Social World

Benjamin David Pyle

RESPONSE

KNOWLEDGE GENERATION AND UNCERTAINTY IN AN UNPREDICTABLE SOCIAL WORLD[†]

BENJAMIN PYLE[*]

CONTENTS

[†] An invited response to Megan T. Stevenson, *Cause, Effect, and the Structure of the Social World*, 103 B.U. L. REV. 2001 (2023).

[*] Associate Professor of Law, Boston University School of Law.

INTRODUCTION

Professor Megan T. Stevenson's Article, *Cause, Effect, and the Structure of the Social World*, is an incredibly important, deep, and thought-provoking argument explaining what we can learn about fundamental causal relationships when we observe few interventions with long-lasting, cascading consequences.[1] It is a profound reflection on empirical work in the social sciences.

A. *Some of the Major Contributions Within* Cause, Effect, and the Structure of the Social World

The Article argues that we have found few, if any, well-identified policy levers that generate outsized, long-term positive impacts for those impacted by the criminal legal system. It offers several explanations for the lack of randomized control trial ("RCT") evaluations with large, non-mechanical effects, but the critical insight is that the social world is composed of stabilizing forces." To make this argument, it begins with empirical work, documenting that hundreds of careful experiments have studied the criminal legal space. Stevenson argues that RCTs are highly credible research designs, and are the type of evidence we ought to trust the most to identify causal relationships. Relative to other forms of causal empirical work, RCTs are more difficult to manipulate and more likely to be published regardless of their findings. Despite several features making RCTs less biased than other designs, these experiments are still more likely to be published and well-known if they find outsized policy impacts. Yet, even with this potential bias, we see few RCTs generating large, long-lasting improvements with respect to many of the outcomes we care about. Those interventions that initially seem promising have difficulty replicating or scaling.

The Article carefully demarcates the scope of the critique, and much of this Response will be spent discussing the boundaries of its argument. The empirical argument focuses on RCTs.[2] RCTs often focus on relatively small-bore solutions. These interventions tend to be small because implementing an RCT often requires navigating normative and practical constraints restricting the scope of policies researchers can test.[3] This Response explores the idea that the

[1] Megan T. Stevenson, *Cause, Effect, and the Structure of the Social World*, 103 B.U. L. REV. 2001 (2023) [hereinafter Stevenson, *Cause, Effect, and the Structure of the Social World*].

[2] The critique's scope extends to programs that can be identified quasi-experimentally. This inference is done partly by analogy to the more systematic evidence presented regarding RCTs.

[3] This echoes a long-standing critique of the scientific process—when we are restricted to "evidence-based reform," we are searching only where it is easiest to look (or at least where we think we can generate the most credible evidence). ROBERT F. BARSKY, NOAM CHOMSKY: A LIFE OF DISSENT 95 (1998) ("Science is a bit like the joke about the drunk who is looking

interventions we are willing to evaluate with an RCT are constrained by political will, ethics, time, costs, and many other factors, and what these constraints imply for the inferential argument. Constraints on RCTs in social science are common, but we may be exceptionally constrained within criminal legal interventions.[4] Understanding which constraints are binding on our knowledge-generation process is essential for interpreting the Article's evidence, its epistemic versus substantive critiques, and ultimately, our ability to improve criminal legal policy.

B. *A Graphical Framework for Engaging with the Article's Framework for Knowledge Generation*

When assessing empirical arguments, thinking about the underlying data-generating processes is helpful. The data the Article uses is previous RCT findings. These are a function of the nature of both the causal world and the scientific knowledge-generation process.

Professor Stevenson offers many important contributions about the world's underlying causal nature and the social scientific publication process, and this reply is limited to what it can engage with by a word count. So, under the theory that a picture is worth one thousand words, I'm going to use a figure to add some additional scaffolding to the conversation, inspired by Professor Stevenson's excellent discussion of a sliding scale of evidence in Section III.B.2.[5] The Figure below shows one stylized representation of how we might learn about the world from empirical work. Moving up the y-axis represents the credibility someone from the evidence-based policy ("EBP") reform movement places in a study or analysis.[6] The degree of EBP credibility is related to how comparable the group

under a lamppost for a key that he has lost on the other side of the street, because that's where the light is. It has no other choice." (quoting Noam Chomsky)).

[4] For many good reasons, we are unlikely to allow a researcher to randomly subject one subset of the population but not another to the possibility of facing the death penalty. Reforms like prison abolition would include large societal shifts poorly suited to an RCT implementation.

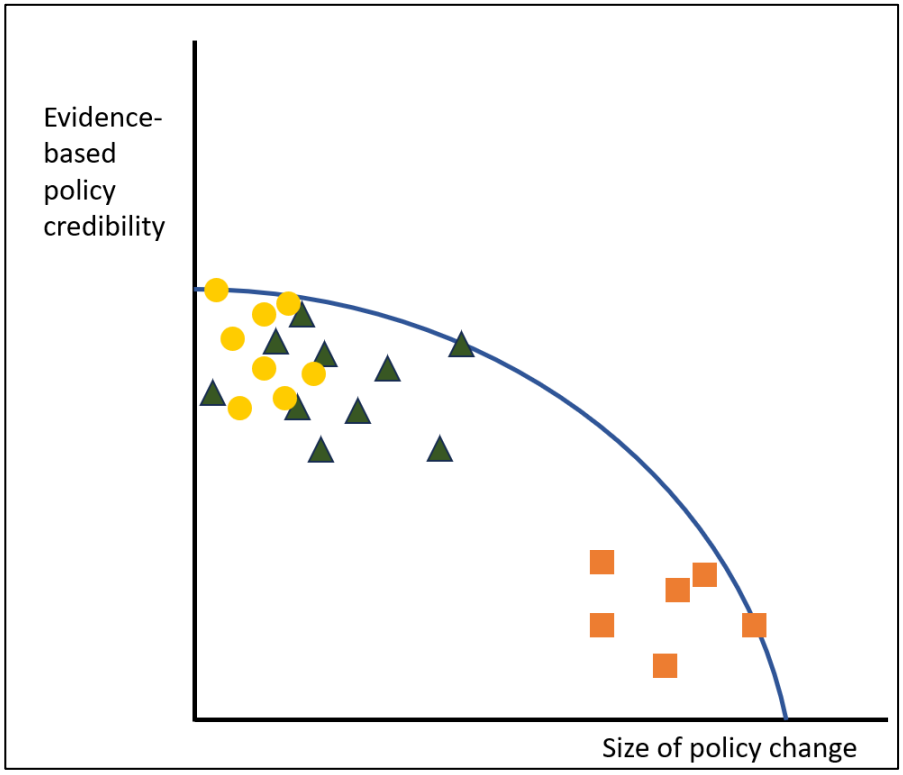
[5] Stevenson, *Cause, Effect, and the Structure of the Social World*, *supra* note 1, at 2037.

[6] Whether EBP reform gets the appropriate weight in policy decisions is a question of some contention. *See Reducing Violence Without Police: A Review of Research Evidence*, JOHN JAY COLL. OF CRIM. J.: RSCH. & EVALUATION CTR. (Nov. 9, 2020), https://johnjayrec.nyc/2020/11/09/av2020/ [https://perma.cc/4D55-Q9KB] ("Policymakers and the public have been told for decades that the 'gold standard' of evaluation evidence is the randomized experiment, or randomized controlled trial (RCT). If all questions relevant for policy and practice in the prevention of violence were amenable to randomized studies, this would be an admirable position. In many areas of social policy, however, some important questions cannot be answered with RCT studies due to logistical, financial, and ethical concerns. This is especially true in the case of violence prevention and violence reduction at the community level. Randomized designs are a valuable resource for providing precise answers to specific questions, but it is also important to ask the right questions and only then select the best method of answering them." (citation omitted)).

receiving the policy intervention is relative to the untreated "counterfactual" group and how convincing the randomization of policy treatment is.[7] Moving to the right on the x-axis represents a larger policy intervention. For instance, to the left might be an evaluation of a policy giving a population $5,000 over three years, while to the right might be prison abolition.

Different research designs tend to fall in different locations in this plane. Yellow dots represent what we have learned from RCTs. Green triangles represent quasi-experimental studies. Orange squares represent cross-country comparisons or comparisons between nonrandomly assigned jurisdictions or people facing very different sets of criminal legal policies.

Figure 1. Evidenced-Based Policy Credibility v. Size of Policy Change



The blue curve represents the knowledge production possibilities frontier ("KPPF"), indicating the most we can learn about the causal world (at least from an EBP perspective) subject to our current abilities and constraints. The Figure

[7] JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MASTERING 'METRICS: THE PATH FROM CAUSE TO EFFECT ch. 1 (2015); Stevenson, *Cause, Effect, and the Structure of the Social World*, *supra* note 1, at 2037.

presents one possibility for what the shape of the frontier might look like, but it is important to highlight that this is a stylized model and not an empirical observation. A critical feature of my depiction is that the KPPF is downward sloping—as we seek to evaluate larger departures from the status quo, we often have lower EBP credibility. While there is some overlap between methodologies, RCTs tend to study smaller policy changes that EBP practitioners place heightened credibility on; quasi-experimental designs include somewhat larger policy interventions but are potentially less credible for the reasons Professor Stevenson argues.[8] Nonrandomized studies often document large differences in outcomes across jurisdictions, with substantially different sets of policies, but are usually deemed less credible causal evidence for any particular policy by the EBP inclined.[9]

The best studies we can currently produce will fall on the edge of the KPPF. Not every study will be on the production frontier, as some RCTs may be less well executed and thus less credible, some quasi-randomization arguments may require assumptions unlikely to be true, and some measurements across different jurisdictions may be done more carefully than others. As barriers to research are removed, we can shift the frontier. For instance, recent advances in statistical methodologies and improved data holdings have vastly increased our ability to conduct quasi-experimental studies tying criminal legal events to a host of previously understudied outcomes like education, family formation, lifetime earnings, and more.[10] Similarly, as more agencies become amenable to working

[8] Professor Stevenson's critique of knowledge generated from quasi-experiments might suggest the KPPF should be a straight line, or even convex, rather than the concave relationship depicted above. The depiction is an attempt to characterize the view of many, but not all, EBP producers and consumers.

[9] It is worth emphasizing here that I have not exhausted the universe of research methodologies and I am not taking a stand myself on what design or research approaches have the most merit, but rather attempting to reflect the various weights EBP places on various approaches in policy evaluation. Descriptive work can also be useful and important evidence for those in the EBP camp for deciding which areas to focus on. EBP would still place a high value in documenting the fact that the X number of people are currently in a certain situation to help prioritize where to structure their search for impactful policies.

Some have argued that economics has moved away from asking bigger questions by placing a higher emphasis on the causal credibility of the research design over time. *See, e.g.*, Shawn Donnan, *A Nobel Laureate Offers a Biting Critique of Economics*, BLOOMBERG (Sept. 29, 2023, 10:23 AM), https://www.bloomberg.com/news/articles/2023-09-29/angus-deaton-s-new-book-says-economists-value-markets-over-people#xj4y7vzkg ("What's known as the 'credibility revolution' in economics in recent decades has seen a focus on real-world studies that have brought a flood of new data and ought to be helping find solutions. But Deaton thinks it has led the profession away from pondering the big questions to focusing on easily quantifiable ones. 'You're finding out very credible results about things you're not very interested in . . . .'").

[10] *See, e.g.*, Keith Finlay, Michael Mueller-Smith & Jordan Papp, *The Criminal Justice Administrative Records System: A Next-Generation Research Data Platform*, SCI. DATA

with researchers and as more money flows into organizations conducting RCTs, the set of policies that can be evaluated increases.

There are reasons to believe the KPPF is expanding. For instance, we will continue to learn important noncausal facts about the criminal legal system (e.g., how many children grew up in families with criminal legal involvement; how many people have records). This progress may expand our ability to evaluate policies both with RCTs (increased funding and resources, more buy-in from institutional actors) and quasi-experimentally (better data measuring more outcomes). One question explored in this Response is how one might think about Professor Stevenson's critique in the context of an expanding KPPF.[11]

THE BOUNDARIES OF THE CRITIQUE

A. *Can We Rule Out Weaker Versions of the Engineer's Worldview?*

Stevenson's major inferential move is to interpret the lack of credible, positive findings (in the sense that most small interventions do not lead to large and lasting change) in the past fifty-plus years as strong evidence that the engineer's view of the world is unlikely to be true. Rather, she argues this empirical pattern is consistent with a world relatively immutable to small policy changes because it is full of stabilizing forces.[12] While the argument does not rely on a formal statistical hypothesis, it may be productive to consider what one might look like. The null hypothesis is that the world is full of stabilizing forces. It is evaluated using evidence from RCTs. If the hypothesis is false, we would expect to see many positive and significant RCTs. The Article covers evidence from hundreds of RCTs, most of which fail to generate long-lasting, cascading impacts. We thus fail to reject the null hypothesis. The argument makes the additional inference that, based on this evidence, we should expect few (if any) interventions to work within the range of all possible interventions of the size that can be evaluated with RCTs.[13]

(Sept. 2022), https://www.nature.com/articles/s41597-022-01620-y [https://perma.cc/2QXL-YYRS].

[11] For instance, the current paradigms and approaches we use to map the social world may not persist forever, and could be replaced with other, better, evidence-based approaches. *See generally* THOMAS KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS (1962).

[12] "Under the engineer's view, social processes are structured and manipulable. RCTs and other causal inference methods are used to map the functioning of the machine, to see what impact a particular lever has." Stevenson, *Cause, Effect, and the Structure of the Social World*, *supra* note 1, at 2004. "Under the engineer's view, the causal structure of the social world can be mapped using RCTs and other scientific methods, and, once mapped, it can be manipulated to achieve social goals. Certain interventions yield such consistent and replicable success that they can be labeled 'best practices.' And meaningful reform can be achieved with reduced risk and uncertainty because the interventions have been rigorously evaluated before scaling up." *Id.* at 2038.

[13] While the evidence Professor Stevenson provides relates to RCTs, she argues that the critique applies to high-quality causal inference generally: "The scope of my claim is thus not

While Professor Stevenson's argument is empirical, our usual formal tools for evaluating such a hypothesis are difficult to implement for large questions like "what is the nature of the social world?" and so we are unable to avail ourselves of some of the clarity and the quantification of uncertainty available in settings where we can formalize the empirical question.[14] That is not to say that Professor Stevenson's argument is not at the appropriate level of formalization for the question at hand—it clearly is—but I hope to reiterate some of the inherent tradeoffs in answering big questions with empirical arguments.

The Article provides convincing evidence that well-understood, easily implementable, and scalable (practically and politically) RCT-style interventions that dramatically improve outcomes in the criminal legal system are scarce compared to the number of attempted policies given our current set of policies, political constraints, and technologies. However, the last statement contained many qualifying adjectives, and I think the evidence is less clear-cut as the various qualifiers are removed. In the remainder of this Response, I hope to continue highlighting some of the many contributions of Professor Stevenson's argument and explore some assumptions dictating the boundaries of the critique. I will do this by exploring why some of these qualifications are important.

My remaining questions can be divided into two loose conceptual buckets. First, how certain are we that we have fully explored the portion of the KPPF that we can currently evaluate with RCTs, and relatedly, how much confidence should we have in rejecting the engineer's view?[15] A related way of framing questions of this type is: If the bulk of the empirical evidence shows that tested interventions have close to zero long-term impact, how tight of a null result is it over the set of all currently possible RCT-measurable policies? Second, if we relax some of the many constraints on what we can test with RCTs (expand the KPPF and move some of the yellow dots upwards and to the right), will we find cascade-generating interventions? Which constraints are the most binding?

A question lurking beneath both points of inquiry is how to precisely characterize the engineer's view. Within the current framing, it is uncertain how many policies need to "work" for the engineer's view to be true. Is thinking of the "engineer's" view as binary or a sliding scale more helpful in understanding some features of the world but not others? Suppose some of the currently more promising programs do replicate and scale. For instance, at the end of Section II,

just limited to interventions evaluable via RCT, it's limited to interventions evaluable via rigorous method of empirical causal inference." *Id.* at 137. There are more examples of studies finding cascades within quasi-experimental work, but given potentially higher rates of publication bias this evidence is harder to evaluate.

[14] Recall the KPPF discussed earlier. This is a big question and would likely fall far to the right.

[15] If we were going to strain the statistical analogy, under the null hypothesis that the engineer's view of the world is true, how surprised should we be that we see few successes? What share of policies does the engineering view think generates cascades? Should this happen less than 5% of the time? Less than 1%?

Professor Stevenson presents evidence of the most promising interventions. Summer jobs potentially have some moderate impacts on crime beyond the simple "direct" incapacitation effect of not committing crimes while busy with a job.[16] Another area of promising RCT-evaluated interventions, at least at the early stages of promise, includes investments in environmental public goods. While additional replication and expansion is needed, restoring blighted houses and installing lighting do, at least for now, seem to be long-lasting, relatively cheap interventions that improve outcomes related to the criminal legal system.[17] IRS auditing enforcement may have long-lasting impacts on tax collection and evasion.[18] Several interventions in early childhood investments in subsidized preschool or cognitive behavioral therapy ("CBT") combined with other interventions have also seemed promising.[19] If all of these interventions turn out

[16] As discussed in the Article, there is also some mixed evidence that social service-based strategies such as summer jobs for disadvantaged youth have some returns. *See generally* Sara B. Heller, *Summer Jobs Reduce Violence Among Disadvantaged Youth*, 346 SCI. 1219 (2014); Jonathan M.V. Davis & Sara B. Heller, *Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs*, 102 REV. ECON. & STAT. 664 (2020).

[17] *See, e.g.*, Aaron Chalfin, Benjamin Hansen, Jason Lerner & Lucie Parker, *Reducing Crime Through Environmental Design: Evidence from a Randomized Experiment of Street Lighting in New York City*, 38 J. QUANTITATIVE CRIMINOLOGY 127, 151 (2022) (lighting and place); David Mitre-Becerril, Sarah Tahamont, Jason Lerner & Aaron Chalfin, *Can Deterrence Persist? Long-Term Evidence from a Randomized Experiment in Street Lighting*, 21 CRIMINOLOGY & PUB. POL'Y 865, 872-75 (2022) (longer-term follow up). *See generally* Ruth Moyer, John M. MacDonald, Greg Ridgeway & Charles C. Branas, *Effect of Remediating Blighted Vacant Land on Shootings: A Citywide Cluster Randomized Trial*, 109 AM. J. PUB. HEALTH 140 (2019) (renovating or cleaning up vacant lots); Charles C. Branas et al., *A Difference-in-Differences Analysis of Health, Safety, and Greening Vacant Urban Space*, 174 AM. J. EPIDEMIOLOGY 1296 (2011) (same); Charles C. Branas et al., *Citywide Cluster Randomized Trial To Restore Blighted Vacant Land and Its Effects on Violence, Crime, and Fear*, 115 PNAS 2946 (2018) (same); Michelle Kondo, Bernadette Hohl, SeungHoon Han & Charles Branas, *Effects of Greening and Community Reuse of Vacant Lots on Crime*, 53 URB. STUD. 3279 (2016) (same); Philip J. Cook & John MacDonald, *Public Safety Through Private Action: An Economic Assessment of BIDs*, 121 ECON. J. 445 (2011) (analyzing effect of business improvement districts on crime); Jeffrey D. Morenoff, Robert J. Sampson & Stephen W. Raudenbush, *Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence*, 39 CRIMINOLOGY 517 (2006) (analyzing effect of proximity to homicide on homicide rates); Kees Keizer, Siegwart Lindenberg & Linda Steg, *The Spreading of Disorder*, 322 SCI. 1681 (2008) (analyzing "broken window theory" and effect of disorder on causing more disorder).

[18] *See generally* William C. Boning, Nathaniel Hendren, Ben Sprung-Keyser & Ellen Stuart, *A Welfare Analysis of Tax Audits Across the Income Distribution* (Nat'l Bureau of Econ. Rsch., Working Paper No. 31376, 2023).

[19] *See generally* Sara B. Heller et al., *Thinking, Fast and Slow? Some Field Experiments To Reduce Crime and Dropout in Chicago*, 132 Q.J. ECON. 1 (2017) (suggesting CBT improves decision making, reduces criminal behavior, and increases high school graduation); Chris Blattman, Julian C. Jamison & Margaret Sheridan, *Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia*, 107 AM. ECON. REV.

to have relatively large impacts and replicate in different settings, is the engineer's view of the world redeemed? What if one does? It may be helpful to characterize the critique as stating that those holding the engineer's view believe more interventions will "work" than our current body of evidence supports. However, more work is needed to characterize precisely what beliefs the evidence supports.

B. *Are There "Light Touch" Policies That We Have Not Yet Considered Achievable Within Current Research Constraints, and if So, Why Not? What Quality Should We Expect Untested Interventions To Be?*

It is helpful to think about the set of reforms that can (currently) be tried and studied within traditional RCT frameworks. In the earlier Figure, they all fall towards the left of the "size of policy change" axis. Still, the set of potential smaller-scale reforms is very large. There are countless ways the criminal legal space could be changed. Most of them have yet to be tried. Still more have yet to be evaluated with a high-quality RCT (or any attempt at causal inference). The fact that we have not found many (or maybe any) replicable cascade-generating interventions is consistent with several potential realities. Two such possibilities are that (1) our current technology and understanding of the world do not allow us to identify promising reforms well, and the share of large-impact, light-touch interventions is small relative to all possible RCT-measurable interventions; and (2) there are very few, if any, light-touch interventions that have large effects.

Suppose we are randomly sampling from the set of policy reforms, something that is likely implied by possibility (1). In that case, the evidence suggests we should reject the hypothesis that most reforms have long-lasting, highly beneficial returns. If we are untargeted in picking policies (or bad at picking) and half of all possible interventions worked, it would be quite surprising that we have not found many interventions that work well. But can we reject that 1 out of every 100 reforms would have the types of effects we are looking for? What about 1 out of every 1000? Answering these types of questions is harder. The answers to these questions dictate whether we reject something one might

1165 (2017) (RCT in Liberia recruiting high-risk men aged eighteen to thirty-five finding CBT reduces antisocial behavior, but only in longer term when combined with monetary grant); Christopher Blattman, Sebastian Chaskel, Julian C. Jamison & Margaret Sheridan, *Cognitive Behavior Therapy Reduces Crime and Violence over 10 Years: Experimental Evidence* (Nat'l Bureau of Econ. Rsch., Working Paper No. 30049, 2023). There is some speculative evidence indicating potentially some benefit from restorative justice interventions in some settings. *See* Yotam Shem-Tov, Steven Raphael & Alissa Skog, *Can Restorative Justice Conferencing Reduce Recidivism? Evidence From the Make-it-Right Program* 16-23 (Nat'l Bureau of Econ. Rsch., Working Paper No. 29150, 2022). There are other potentially promising, but by no means proven or cascading, programs evaluated with RCT. *See Search Rated Programs*, NAT'L INST. OF JUST., https://crimesolutions.ojp.gov/rated-programs (last visited Nov. 20, 2023).

call the modest engineer's view—that some interventions in the world can generate outsized improvement, but they are rare and hard to find.

Professor Stevenson's critique is most potent if we are in the second world, where we can successfully implement and evaluate the most promising programs and can identify the best programs with reasonable ex ante certainty. In this case, we should be less hopeful that any RCT-evaluable reforms will generate cascades. If we have successfully identified and replicated five hundred of the most promising programs and all failed to yield the hoped-for results, it is unlikely that any will. Professor Stevenson questions, "if research paradigms are so resistant to the knowledge that *they themselves generate*, how can we be confident in our systems of knowledge generation?"[20] How confident are we that we are testing the most promising policies?[21] How skilled we are at identifying and testing promising programs is an important assumption necessary to draw evidence from existing programs to yet untried or untested programs.

It is unclear whether the current evidence allows us to rule out these various possibilities. And it matters which world we are in, both for our understanding of how the social world works and for how we impact policy. If barriers to finding these policies are practical, more resources can solve the problem, but we should be more cautious if more fundamental constraints are binding our search.

Is there evidence regarding how well we choose programs to evaluate? Given the level of abstraction in the above argument, it might be helpful to fix ideas in a specific example. Without debating how much weight we should place on causal evidence produced by quasi-experimental evidence (Professor Stevenson convincingly raises several concerns with the production of this type of knowledge), it may be productive to consider whether there is any RCT evidence regarding several interventions considered promising in these types of studies. Concretely, should the finding that RCTs on job-training programs, policing "hot spots," and several other criminal legal programs have had limited results suggest that an untested (at least by RCT) policy of lead abatement will have limited effects? This Response has argued that the strength of inference regarding the underlying nature of the world might depend upon *why* we have tested job-training programs but not lead abatement. Does the fact that we have not tested this policy tell us something about which policies we evaluate?

Many EBP advocates view lead abatement as a potentially efficacious policy.[22] However, a review of the literature reveals little RCT evidence on lead

[20] Stevenson, *Cause, Effect, and the Structure of the Social World*, *supra* note 1, at 2046.

[21] There is an interesting tension in thinking that we are both pretty good engineers in identifying the most promising levers to pull, and that the levers seem to do a lot less than hoped for.

[22] The above focused on lead abatement, but there are other promising interventions that have limited RCT evidence. For instance, diversion from the criminal legal system prior to entry. *See generally* Elsa Augustine, Johanna Lacoe, Steven Raphael & Alissa Skog, *The Impact of Felony Diversion in San Francisco*, 41 J. POL'Y ANALYSIS & MGMT. 683 (2022);

exposure and abatement (or exposure to other environmental pollutants).[23] Is this because such RCTs are deemed not promising? Or are there other explanations? For instance, one prominent RCT in this space (not studying crime as an outcome) raised severe ethical and equity concerns.[24] Indeed, interventions focusing on reducing exposure to lead somewhat straddle criminal justice and public health, and are thus potentially the most likely criminal legal interventions to fall outside of Professor Stevenson's critique (and, at least based on theoretical and extant quasi-experimental evidence, have the potential to produce cascades).[25]

An RCT that randomly removes lead from areas with old and deteriorating paint and replaces old windows and doors seems theoretically feasible. These policies are relatively small bore and are likely reasonably comparable in cost to other programs explored by RCTs. If the engineer's view of the world is false, should we expect a well-executed RCT on these policies to generate small or null results?

If we are in the scenario where most, if not all, interventions we have tried fail to generate cascades, building certainty as to the reason why is important. It might be that the causal world is a hopelessly complex ecosystem that can never be mapped. It might also be that actors within the system work to preserve the

Michael Mueller-Smith & Kevin T. Schnepel, *Diversion in the Criminal Justice System*, 88 REV. ECON. STUD. 883 (2021); Amanda Y. Agan, Jennifer L. Doleac & Anna Harvey, *Misdemeanor Prosecution*, 138 Q.J. ECON. 1453 (2023).

[23] Maria Jose Talayero, C. Rebecca Robbins, Emily R. Smith & Carlos Santos-Burgoa, *The Association Between Lead Exposure and Crime: A Systematic Review*, 3 PLOS GLOB. PUB. HEALTH, Aug. 1, 2023, at 1, 14-17; *see also Episode 16: Stephen Billings*, PROBABLE CAUSATION, at 16:00 (Nov. 12, 2019), https://www.probablecausation.com/podcasts/episode-16-stephen-billings (discussing limitations to running RCTs in lead abatement area); Evan Herrnstadt, Anthony Heyes, Erich Muehlegger & Soodeh Saberian, *Air Pollution and Criminal Activity: Microgeographic Evidence from Chicago*, 13 AM. ECON. J.: APPLIED ECON. 70, 76-81 (2021); Ryan W. Allen, Prabjit K. Barn & Bruce P. Lanphear, *Randomized Controlled Trials in Environmental Health Research: Unethical or Underutilized?*, 12 PLOS MEDICINE, Jan. 2015, at 1, 1-2 (2015) (discussing why there are limited RCTs in this area).

[24] *See generally* David R. Buchanan & Franklin G. Miller, *Justice and Fairness in the Kennedy Krieger Institute Lead Paint Study: The Ethics of Public Health Research on Less Expensive, Less Effective Interventions*, 96 AM. J. PUB. HEALTH 781 (2006).

[25] Stevenson exempts public health from her critique. Stevenson, *Cause, Effect, and the Structure of the Social World*, *supra* note 1, at 2033. The divisions between fields here may be somewhat malleable, as at least some in the public health space would include violent behavior as an appropriate object of study. *See, e.g.*, LINDA L. DAHLBERG & JAMES A. MERCY, CTR. FOR DISEASE CONTROL & PREVENTION, THE HISTORY OF VIOLENCE AS A PUBLIC HEALTH ISSUE 1 (2009) ("Violence is now clearly recognized as a public health problem . . . ."). If lead interventions are public health, should CBT also be considered? What about changing punishments for juveniles to reflect physical brain development? Providing food or diet interventions? Perhaps the dividing line is targeting health and the physical body. But even with this guiding principle in mind, it may be difficult to draw lines around what should and should not be exempted from the critique.

status quo, restricting experiments away from those that might be most promising or are working (intentionally or not) against certain outcomes. One reason this might arise is political constraints. While researchers have incentives to find "cascade" policies, it is less evident that other institutional actors who need to participate in many RCTs have similar motivations. If all the levers that would be most effective require buy-in from the groups controlling the system, and those controlling the system prefer the status quo, we might expect to be unable to test the most promising levers. One critical view of the criminal legal system is that it is a social system designed to exert power to maintain a particular societal order. Some have critiqued evidence-based policy research as requiring at least some buy-in from the system's actors. Suppose political constraints are the main reason past RCTs have been unsuccessful. In that case, we are in a world of practical barriers to the engineer's goal of mapping out causal pathways rather than fundamentally unmappable or unchangeable reality.[26]

C. *Can We Currently, or Ever, Convincingly Causally Map Larger Interventions? Do These Have a Better Chance of Generating Cascades? How Much (and How Quickly) Can We Expand the Knowledge Production Possibilities Frontier?*

RCTs have many constraints governing what policies can be tested. These constraints might change, allowing assessment of larger, more complex interventions.

One constraint might be resources. For instance, the size of the question we are evaluating in an RCT is related to how large of an intervention we can fund. More minor interventions are easier to implement and measure. For example, giving the randomly treated population $100 is far cheaper than providing them each a house. This problem compounds if we think cascading impact requires pulling the right combination of levers together simultaneously. While we have explored some smaller-scale interventions with RCTs, we are not close to evaluating the complete set of potentially promising, more resource-intensive interventions. Is the lesson to learn from small-scale intervention's modest impact that we should go bigger or not at all? We have likely not hit a fundamental constraint on the size of intervention that can be credibly evaluated with an RCT.

Another constraint might be ethics. There are good reasons to limit RCTs, especially in the criminal legal space. Certain policies will not be randomly assigned and measured for horizontal equity and fairness. While questions like the death penalty and sentence length might be up for debate in the broad policy change space, researchers' random manipulation of these policies is off the table.

[26] In practice these "realities" might end up being the same, depending on how mutable the political constraints are. It may be possible for the engineer's *substantive* view to be correct, even if the *epistemological* view is not.

Beyond cost, political buy-in, and ethics, there is potentially a more fundamental upper limit to the size of the intervention that can be considered by an RCT (or quasi-experimental method, for that matter). One challenge with scaling up interventions, even if we have the resources and political will, is finding a reasonable comparison group. If it is the case that the only convincing comparison group is people who are at least somewhat connected to the treated population, and the only treatments that will have cascading effects are those that dramatically change the environment for the treated people in such a way that the treatment also reaches those from the comparison group, we will have difficulties in identifying causal effects. That is, it might be that even with infinite resources and will, credibly holding a roughly comparable group as untreated may be implausible for interventions of a certain large scale. As Professor Stevenson points out, this is inherently a small "c" conservative approach. The critique is most potent if this fundamental constraint is binding. While we might imagine scenarios where we can expand the KPPF by increasing funding, data, and political access, our ability to identify comparable counterfactual groups likely faces an immutable, natural upper bound.

We cannot rule out a world where most cascading changes require an immense impetus with small-bore RCT evidence. If we have reached the natural upper bound of the size of intervention we can credibly evaluate and are selecting the most promising policies as best as we (ever) can, it is unlikely we will ever find outsized policy levers. If other, more malleable constraints are binding, perhaps one day we will.

CONCLUSION

An implication of Professor Stevenson's argument is that agents are in a sticky, local equilibrium.[27] This feature is well illustrated by the orange in a bowl

[27] One question this Response cannot fully engage with, but is mentioned in Professor Stevenson's conclusion is the relationship between sticky equilibrium and whether agents are optimizing. The Article ends by making a case for optimism and one that requires some additional assumptions. Professor Stevenson interprets the evidence presented as suggesting that "people had maximized their utility subject to constraints" and that the easy improvements have been made, such that "any barriers to success that were readily moveable had already been moved." Stevenson, *Cause, Effect, and the Structure of the Social World*, *supra* note 1, at 2047. The argument here is, roughly, if a small investment could have been made to improve a person's life from their given circumstances dramatically, it has already been made thanks to the individual's (or community's) efforts. It is optimistic in the sense that individual people (and perhaps communities) are doing the best they can, at least given the constraints they face. It is somewhat more pessimistic if you view the current equilibrium as untenable (as I think many might upon a close examination of the criminal legal system), as it suggests that there isn't an easy path forward. It also, perhaps, relies on an assumption that people are optimizing. As a person trained in economics, I am not unsympathetic to this assumption. However, someone who does not buy into this optimization argument might view it more pessimistically as individual self-destructive habits are persistent over time and difficult to break.

exposition. While the orange may be pushed up the side of the bowl by various interventions, it eventually returns to an equilibrium at the bottom of the bowl. Evidence from RCTs can be mapped to relatively modest pushes of the orange within the bowl. But perhaps the sides of the bowl are not infinitely tall. A large enough push could move the orange out of the bowl and to a different equilibrium resting place. We do not know how high the bowl walls are or if we are in an orange-bowl world. The tide and wind analogy suggests that there are not multiple equilibria and that regardless of the size of our intervention, we are unlikely to have a long-lasting impact.[28] Evidence that interventions have had little long-term impact is consistent with worlds that follow either causal framework. Nor are these analogies binaries, as both could accurately describe parts of the world.[29]

Another way of approaching this question is to think about what model of the world consists of discrete, limited-scope interventions that do have large or replicable impacts. If that model were the case, we would expect to be living in a world full of non-linear returns, multiple equilibrium, or thresholds. We can rule out that the interventions we have tried consistently generate jumps from one equilibrium to another. With some inductive assumptions, we might think all small policy interventions are unlikely to be dramatically life-changing. As discussed earlier, this might be because the interventions have been too small or poorly targeted to overcome whatever threshold has to be met to move people from one way of life to another. It could also be that few interventions do.[30]

[28] Monica P. Bhatt, Jonathan Guryan, Jens Ludwig & Anuj K. Shah, *Scope Challenges to Social Impact* (Nat'l Bureau of Econ. Rsch., Working Paper No. 28406, 2021). might help us understand what sorts of interventions are likely to have larger "scope." First, we are more likely to have lasting impacts on behaviors that rely on fewer decisions, all else equal (e.g., setting up a savings account). Interventions that impact a series of decisions made in the same context. Interventions that have high habit-formation. Interventions that target multiple decisions motivated by a shared reason. If we see most interventions having little scope, this suggests that most decisions rely on many decisions that policy makers have difficulty reaching simultaneously, perhaps due to the decisions relying on many disparate reasons. It also suggests that breaking previous habits and establishing new habits is rare.

[29] Some outcomes may have many factors driving the choices, while others may depend upon fewer features of the world. Similarly, each of these pathways may require differing degrees of intervention to shape end results.

[30] As Professor Stevenson points out, that doesn't mean that these programs aren't having important direct effects. Measuring the direct effects of these programs is also likely a worthy cause even without cascading returns (some programs may not even achieve their intended direct effects)! Interventions alleviating hunger, providing short-term jobs, or providing medical care all may have positive impacts that are worthy pursuits. *See generally* Alissa Fishbane, Aurelie Ouss & Anuj K. Shah, *Behavioral Nudges Reduce Failure To Appear for Court*, 370 SCI. 682 (2020); Emily Owens & CarlyWill Sloan, *Can Text Message Reduce Incarceration in Rural and Vulnerable Populations*, 42 J. POL'Y ANALYSIS & MGMT. 992 (2023) (information intervention reducing failures to appear and incarceration); Zoë Cullen, Will Dobbie & Mitchell Hoffman, *Increasing the Demand for Workers with a Criminal*

Cause, Effect, and the Structure of the Social World is an essential piece that raises fundamental questions about our understanding of the world and the policies we consider and measure. It highlights compelling reasons to be skeptical that we will uncover many, if any, generalizable and replicable policy interventions that have large, lasting effects on people's lives: that the underlying nature of the world is complex, evolving, and full of stabilizing forces. What we currently know about what works is modest. The Article cautions that the nature of the world is such that we may never find a moderate intervention that generates outsized results. What is clear from Professor Stevenson's Article is that those holding an ambitious engineer's view of the world should update their beliefs given the evidence, and transparency and modesty in the scientific process should be valued. This Response has highlighted several questions that we should seek to answer to help us interpret the boundaries and implications of this critique.

Record, 138 Q.J. ECON. 103 (2023) (discussing several policies increasing hiring for workers with criminal records).