Columbia Law School

Scholarship Archive

2019

Broken Experimentation, Sham Evidence-Based Policy

Kristen Underhill
Columbia Law School

Recommended Citation

YALE LAW & POLICY REVIEW

Broken Experimentation, Sham Evidence-Based Policy

Kristen Underhill*

Evidence-based policy is gaining attention, and legislation and agency regulation have been no exception to calls for greater uptake of research evidence. Indeed, current interest in "moneyball for government" is part of a long history of efforts to promote research-based decisions in government, from the U.S. Census to cost-benefit analysis. But although evidence-based policy-making (EBPM) is often both feasible and desirable, there are reasons to be skeptical of the capacity of EBPM in governmental decision-making. EBPM is itself bounded by limits on rationality, the capacity of science, the objectivity of science, and the authority we wish to give technocrats. Where values are highly contentious, efforts to produce and use evidence in legislative and regulatory decisions may go so far awry that they become "sham" versions of evidence-based choices. In this Article, I name several of these sham practices, including the distortion of evidence, the engagement in "terminal" experimentation that destabilizes governmental programs, and "ratcheting" actions that defund entire priorities rather than individual approaches. Broken experimentation is also common, with evaluations of government programming and policies neglecting or misusing opportunities to provide rigorous evidence. I argue that the stakes of these misuses are high, resulting in losses of welfare and efficiency, erosion of scientific legitimacy, and infringement on the dignity of human subjects. But where genuine engagement with empirical evidence is possible, the game is surely worth the candle. This Article proposes novel ways to promote responsible uses of empirical evidence in both legislation and agency regulation, including evaluation mandates, pre-registration of evaluation protocols and transparency of research reports, ex ante decision commitments, and more attentive uses of judicial review.

INTRODUCTION

    In 2012, death penalty supporters and detractors eagerly awaited the publication of a report by the National Research Council (NRC), two years in the writing, that sought to resolve the question of whether the penalty in

fact deters homicide.[1] The NRC's prior effort to do so in 1978 had been inconclusive, finding that existing studies "provide[d] no useful evidence on the deterrent effect of capital punishment."[2] In the ensuing decades, however, a mass of new research had emerged with wildly divergent findings, ranging from showings of extreme deterrence to the suggestion of increased homicides where the death penalty is imposed.[3] Rancorous disagreements flourished in both politics and the scientific literature. The new committee would weigh this body of evidence, appraise its methodology, and issue an updated conclusion. But when the report emerged, researchers and policy advocates alike were bewildered. Despite all of the intervening research, the committee had concluded that "all of the research about deterrence and the death penalty done in the past generation, including by some first-rank scholars at the most prestigious universities, should be ignored."[4] Evidence, therefore, could be no guide to state decisions on whether to retain or abolish the penalty—states were, as they long had been, left to resolve the question primarily on the basis of values rather than empirical evidence.

Frustration about the weight and direction of empirical evidence across a range of issues—the deterrence effect of the death penalty,[5] the extent to

---

1.   *See* NAT'L RESEARCH COUNCIL, DETERRENCE AND THE DEATH PENALTY (Daniel S. Nagin & John V. Pepper, eds., 2012). I am grateful to Josh Kleinfeld for suggesting this case study.

2.   *Id*. at 1 (quoting the NRC report from 1978).

3.   *Id.* at 1; *see also* John J. Donohue III & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791, 793 (2006).

4.   Editorial Board, *The Myth of Deterrence*, N.Y. TIMES (Apr. 27, 2012), https://www.nytimes.com/2012/04/28/opinion/the-myth-of-deterrence.html [https://perma.cc/JWA9-79C5].

5.   NAT'L RESEARCH COUNCIL, *supra* note 1.

which sanctions[6] and walls[7] deter migration, the extent to which supervised injection facilities reduce opioid overdoses,[8] to name a few—has taken center stage because, by and large, we think evidence should matter in legislation and agency regulation. Indeed, evidence-based policy-making (EBPM) may be having a popular moment. EBPM refers to the systematic use of empirical research evidence by legislators and agency personnel when making government decisions, and it has long been the province of policy wonks and scholars eager for research evidence to play a larger role in decision-making. The process of EBPM entails both the generation and use of empirical evidence at several time points, suggesting that regulators and legislators should (1) consult past research evidence when making policy choices, including evidence that characterizes a public problem and the likely impacts of different policy options for solving it; (2) when making a public policy decision, fund or allow research that will evaluate the impacts of that decision; and (3) after evaluation results are available, reassess and adjust their policy decisions in light of the new evidence.

In recent years, the House and Senate have both taken an interest in the promotion of EBPM, launching a federal Commission on Evidence-Based Policymaking to make recommendations promoting evidence-based practices in the federal government—and, more recently, enacting those recommendations as law early in 2019.[9] These efforts come on the heels of the Obama presidency, the most hospitable administration yet for empirical research in federal regulation, and efforts to rebrand EBPM as "moneyball for government" are gaining popular cache. Observers of the Trump administration's reshaping and retraction of Obama-era scientific committees, rejection of "evidence-based" terminology, and efforts to redefine "evidence" are galvanized to point out affronts to empirical

---

6. *See, e.g.*, Samuel Bazzi et al., *Deterring Illegal Entry: Migrant Sanctions and Recidivism in Border Apprehensions* (Nat'l Bureau of Econ. Research, Working Paper No. 25100, 1981), https://www.nber.org/papers/w25100 [https://perma.cc/TVV7-CKMJ].

7. Gretchen Frazee, *Trump Says Walls Work. It's Much More Complicated*, PBS (Jan. 9, 2019), https://www.pbs.org/newshour/nation/trump-says-walls-work-its-much-more-complicated [https://perma.cc/YLS7-5RGA].

8. Elana Gordon, *What's the Evidence That Supervised Injection Sites Save Lives?*, NPR (Sept. 7, 2018, 2:40 PM), https://www.npr.org/sections/health-shots/2018/09/07/645609248/whats-the-evidence-that-supervised-drug-injection-sites-save-lives [https://perma.cc/Z2BJ-Q23Q].

9. Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2019).

research. Grassroots efforts to bring data to bear on political issues are also ubiquitous.

Against this backdrop, however, questionable uses of evidence and evidence-based policy-making run rampant. High-quality research on policy decisions is often absent, and evaluation mandates are unfunded or toothless, culminating in research that is poorly designed or irrelevant to policy choices. Some policy research is forcibly stalled—as seen in a 2017 congressional decision to bar all randomized trials of school voucher programs after one such trial showed a lack of benefit.[10] "Broken experimentation" and a lack of rigorous policy evaluation provides evidence that is misleading at best, depriving decision-makers and those who implement governmental programs of the evidence needed to improve program designs.

Pathological uses of *existing* research evidence are similarly ubiquitous. In the last few years, for example, efforts to redefine evidence that may be considered by EPA scientific committees, as well as experts eligible for input on such committees, are reshaping the nature of the evidence considered by the agency.[11] Multiple categories of programming have been canceled based on evaluations of individual programs within a larger group, such as the recent revocation of after-school programming[12] and teen pregnancy prevention funds.[13] And claims to "evidence-based" authority for the

---

10. Sarah D. Sparks, *Congress Budget Deal Bans New Gold-Standard Studies of Federal Vouchers*, EDUC. WEEK: BLOGS (May 2, 2017, 9:44 AM), https://blogs.edweek.org/edweek/inside-school-research/2017/05/congress_budget_deal_bans_new.html [https://perma.cc/5Q9N-SSQK]; *see also* Adam Gamoran, *Evidence-Based Policy in the Real World: A Cautionary View*, 678 ANNALS AMERICAN ACAD. POL. & SOC. SCI. (AAPSS) 180, 185 (2018) (describing the bar on randomized trials as "simply banish[ing] unwanted evidence").

11. *See infra* notes 179-180 and accompanying text.

12. Robert Gordon & Ron Haskins, *The Trump Administration's Misleading Embrace of 'Evidence,'* POLITICO (Mar. 31, 2017, 5:35 AM), https://www.politico.com/agenda/story/2017/03/the-trump-administrations-misleading-embrace-of-evidence-000385 [https://perma.cc/JC8V-TX3Q].

13. Alison Kodjak, *Trump Administration Sued over Ending Funding of Teen Pregnancy Programs*, NPR (Feb. 15, 2018, 11:00 AM), https://www.npr.org/sections/health-shots/2018/02/15/585879601/trump-administration-sued-for-ending-funding-of-teen-pregnancy-programs [https://perma.cc/R2WZ-GMQR].

promotion of initiatives with slim or contrary evidence are pervasive, so much so that the term is easily viewed as an empty claim to authority.

One way to interpret these events is through the lens of political economy. Evidence—including the production of information, presentation and classification of facts, and selection of people recognized as authorities—is one tool among many in decisions that are inevitably a matter of political capital. Mobilization, distortion, and selective production of evidence is to be expected. Another interpretation is as a matter of institutional capacity; legislatures, federal agencies, and the humans who make decisions therein are bounded in their rationality and the resources available to produce and use research evidence. But while I do not dispute either account, this Article offers context drawn from a longer history of research on the capacities and limitations of evidence-based policy. The theoretical development of EBPM stretches back decades, if not centuries, with sophisticated models of how research evidence may be used productively in some, but not all, government decisions. EBPM advocates from several recent generations have been clear-eyed about the limitations of research evidence in practice, but much of this thinking has been absent from the discussion of "evidence-based" ideals in legal scholarship.

My goals in this Article begin with an effort to place current interest in EBPM in context, examining the theoretical appeal and advantages of EBPM, the limitations of evidence and its uptake in government decision-making, and a brief history of the rise and uncertain trajectory of EBPM at the federal level. The second aim of this work is to identify ways in which EBPM can go wrong. Building on recent examples in "evidence-based" efforts, I develop an original taxonomy of specific pathologies, ranging from the neglect and deliberate undermining of research (broken experimentation), to the distortion of evidence (a form of sham EBPM), to the disingenuous use of selective evidence to de-fund entire program categories (ratcheting). A third goal of this Article is to illuminate the stakes, potential costs, and possible benefits of sham EBPM, including implications for welfare, efficiency, scientific legitimacy, the production of knowledge, human subjects enrolled in governmental studies, and potential positive externalities.

Building on these arguments, my final aim herein is to propose remedies for the current pathologies in the production and uses of evidence. I begin, however, with a call to set a more realistic scope for the policy areas where we might hope to introduce research evidence productively. Some areas are more prone to EBPM pathologies than others, particularly problems of evidence distortion, weakening scientific legitimacy, and burdens on human subjects who enroll in research with no chance of policy payoff. But where EBPM is possible—where regulators or legislators are

supportive of efforts to use empirical evidence productively—I propose that innovations such as ex ante policy commitments (e.g., pre-commitments contingent on research findings), evaluation mandates, and rules for evaluation transparency may avoid some of the pitfalls that currently stymie EBPM.

This Article proceeds in the following Parts. Part I begins with the merits and limits of evidence-based policy-making, a process that includes both governmental decision-making and subsequent evaluation of those choices. In its ideal form, EBPM improves the likelihood that governmental actions *will achieve their own ends*, making it agnostic to policy preferences. EBPM asks regulators and legislators to identify, synthesize, and interpret the empirical evidence informing each decision; to make decisions in the direction favored by the evidence; to evaluate the outcomes of decisions; and to update decisions as informed by new evidence. This Part will provide a brief overview of EBPM's inroads into regulation and legislation.

Part II turns to ways in which EBPM can fall so far short of the ideal as to be "sham," with costs for both the legitimacy and the practical consequences of government choices. I focus first on common pathologies in governmental efforts to generate evidence (broken experimentation), and then turn to visible pathologies in the application of evidence (sham EBPM). Here I consider four central pathologies of sham EBPM: technical breakdowns in the uptake of evidence, distortion of evidence, terminal or symbolic experimentation without broader adoption of results, and ratcheting cuts to program categories. The central concerns of this Part are deep-seated methodological flaws in evidence generation, as well as irreparable gaps in the feedback loops between evaluation and policy decisions.

Part III will consider the ethics and costs of broken experimentation and sham EBPM, including consequences for allocation of resources, the legitimacy of science, the interests of individual human subjects, and the communities those subjects represent. I will argue here that, despite some positive externalities, sham EBPM can cause harm, including wasting public resources, misleading the public, eroding scientific legitimacy, and threatening the dignity of human subjects.

In Part IV, I seek to reclaim EBPM and governmental experimentation by identifying conditions where program and policy evaluations are likely to be both rigorous and effective in shaping policy. By extension, I will also identify conditions where evidence is likely to be distorted, misleading, or irrelevant. Evaluation should be both more and less used than it currently is—more so in contexts where it would be effective, but perhaps *less* so when there is reason to believe it will skew or contribute little to decisions. Where regulators and legislators do build evaluation into their policy

choices, they have a range of legal options to avoid sham evaluations—both by setting explicit terms for study design, but also by taking affirmative, ex ante, and perhaps binding steps to define how results will affect subsequent policy choices.

I. THE CASE FOR (BOUNDED) EVIDENCE-BASED POLICY-MAKING

It may seem uncontroversial to argue that legislators and regulators should make use of empirical research evidence in their decisions and then evaluate the subsequent effects. Gathering information about the effects of regulatory choices is a fundamental part of experimentalism[14] and new governance,[15] and ex post evaluations of policy decisions are essential for updating those choices over time. Recent legislation[16] and scholarship[17] aim to build momentum for uses of empirical evidence in policy-making processes, with some renewing hopes that attentiveness to statistical evidence may break partisan stalemates.

But debate about the proper role of empirical evidence in governmental decision-making is wide-ranging and often contentious, replete with both practical and normative challenges. The practice and promotion of EBPM was accelerated by the UK Labour Party's commitment to "what works,"[18] the growing embrace of EBPM in US federal agencies, and the academic rise of "evidence-based" terminology in the 1990s. Like evidence-based practice movements in other fields,[19] EBPM has adopted many principles of evidence-based medicine: "the conscientious, explicit and judicious use of

---

14. *See* Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267 (1998).

15. *See* Orly Lobel, *The Renew Deal: The Fall of Regulation and the Rise of Governance in Contemporary Legal Thought*, 89 MINN. L. REV. 342 (2004); Victoria F. Nourse & Gregory C. Shaffer, *Empiricism, Experimentalism, and Conditional Theory*, 67 SMU L. REV. 141 (2014).

16. COMM'N ON EVIDENCE-BASED POLICYMAKING, THE PROMISE OF EVIDENCE–BASED POLICY MAKING (2017).

17. EVIDENCE-BASED SOCIAL POLICY: THE PROMISE AND CHALLENGES OF A MOVEMENT, 678 ANNALS AAPSS 1 (Ron Haskins ed., 2018).

18. KATHERINE SMITH, BEYOND EVIDENCE-BASED POLICY IN PUBLIC HEALTH 11–13 (2013).

19. *See* LIZ TRINDER & SHIRLEY REYNOLDS, EVIDENCE-BASED PRACTICE: A CRITICAL REAPPRAISAL (2000) (canvassing the diffusion of evidence-based medicine principles through medical and nursing practice, public health, education, social work, criminal justice, and other areas).

current best evidence in making decisions,"[20] with priority given to evidence from methodologically rigorous research designs.[21] Translated to legislation and regulation, EBPM asks decisions-makers to identify and use empirical evidence when making choices, to weigh evidence according to methodological quality, to evaluate the impacts of those decisions, and to modify choices that do not work as planned.

This Part will consider the case for using and generating empirical evidence through policy choices, followed by some of the reasons why actual policy development inevitably diverges from the technocratic ideal. "Bounded" EBPM notwithstanding these limitations, however, is both desirable and feasible. Even if policy based entirely on "what works" is elusive, there is ample room for using empirical data in both legislation and regulation once the goals of those policy choices are defined. This Part will then consider past and recent efforts to promote EBPM in practice.[22]

A. Judging Decisions by Their Own Lights

Evidence-based policy-making, in short, is the use of empirical research findings to inform policy choices.[23] In 2016-2017, Congress convened the federal Commission on Evidence-Based Policymaking to issue

---

20.   David L. Sackett et al., *Commission on Evidence-Based Policy-Making*, 312 BRIT. MED. J. 71 (1996). For a retrospective review of the evidence-based medicine movement, see Trisha Greenhalgh et al., *Evidence-Based Medicine: A Movement in Crisis?*, 348 BRIT. MED. J. g3725 (2014).

21.   Gordan H. Guyatt et al., *Users' Guides to the Medical Literature*, 284 JAMA 1290, 1292-1293 (2000) (providing a "hierarchy of strength of evidence for treatment decisions" based on study design and recommending that physicians "look for the highest available evidence from the hierarchy" when making decisions).

22.   For simplicity, this account glosses over differences between different categories of "policy-makers" (legislators, agency appointees, career civil servants, etc.), different categories of "policy" itself (legislation, agency rules, guidance documents, informal practices, programmatic choices, etc.), and different subject areas (health care, education, etc.). I readily recognize, however, that the capacity and boundaries of EBPM will be context-dependent, and that some areas and policy choices are more amenable to uses of empirical evidence.

23.   PAUL CAIRNEY, THE POLITICS OF EVIDENCE-BASED POLICY MAKING 2 (2016); Ron Haskins, *Evidence-Based Policy: The Movement, the Goals, the Promise*, 678 ANNALS AAPSS 8 (2018); Brian Head, *Reconsidering Evidence-Based Policy: Key Issues and Challenges*, 29 POL'Y & SOC'Y 77, 84 (2010).

recommendations for supporting evidence-based practice in federal policy-making.[24] The Commission describes EBPM as follows:

> The Commission envisions a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy . . . . "[E]vidence" . . . refer[s] to evidence produced by "statistical activities" with a "statistical purpose" that is potentially useful when evaluating government programs and policies . . . . The Commission defines evidence-based policymaking as the application of evidence to inform decisions in government.[25]

EBPM asks policy-makers to engage continuously with past research evidence, as well as to support the creation of new, up-to-date empirical research findings that will inform subsequent policy decisions. To clarify these dynamics, we might follow the cycle beginning with the desire to intervene in a public problem, such as opioid overdose deaths. Legislators and regulators engaging in EBPM would ask research questions of governmental and nongovernmental researchers, focusing on the problem and possible solutions. Many forms of evidence and research methodologies are useful at this stage, and evidence should be fit for purpose.[26] In this example, epidemiological methods could provide information about the current extent, distribution, and potential causes of overdose deaths. Qualitative methods can explore behavioral choices and unanticipated contextual factors that may affect the feasibility of policy solutions. Randomized trials, cohort studies, or methods such as difference-in-difference comparisons or regression discontinuity designs could help to identify the impacts of prior efforts to minimize overdose deaths (e.g., expansions of medication-assisted treatment, prescription limits, establishment of supervised injection facilities). Feasibility studies can identify whether particular policy choices are likely to be acceptable in the current context.

With evidence assembled, policy-makers engaging in EBPM would then review the methodology and strength of findings (or delegate the task to others). At this stage, decision-makers are primarily *consumers* of evidence that was gathered in past studies or generated immediately in the current

---

24. Evidence-Based Policymaking Commission Act of 2016, Pub. L. No. 114-140, 130 Stat. 317 (2016).

25. COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16, at 8, 11.

26. Head, *supra* note 23, at 79, 84 (noting "the need to incorporate a wide range of methodologies and forms of knowledge" in evidence-based practice).

context. Decision-makers then make choices that integrate evidence with other considerations, such as legal constraints, concerns about autonomy or dignity, and the availability of resources.

Once decisions are made, policy-makers' role in EBPM shifts from consumer to *generator* (or perhaps patron) of research evidence. Decision-makers committed to EBPM either allow or provide for robust research that evaluates the public impacts of the policies they have put in place. For example, state regulators may choose to impose prescription limits on opioids and then mandate and fund an independent evaluation of how the rule affects overdose deaths over time. The rigor of program evaluations might benefit from implementation choices, such as rolling out a policy in stages or randomizing geographic regions to treatment or control. Ideally, policy-makers would fund these evaluations as part of implementation, but it serves similar ends to partner with outside researchers and funders.

As up-to-date research findings become available, the role of policy-makers in EBPM shifts back from generating to *consuming* evidence. Decision-makers committed to EBPM regularly review the evidence and revisit their policy choices. Where policies have failed to serve important ends, or where they have caused unintended harms, these decision-makers seek to change or eliminate them. And where policies are modified, EBPM envisions new evaluations of the modifications. EBPM thus describes a continuing cycle of past and future research, and it considers the policy-maker's role as both consulting and generating research findings.

Although EBPM envisions a continuous cycle of research generation and uptake, the stepwise process resonates with linear views of policy-making. Namely, these linear views suggest that policy development is an orderly progression including goal-setting, policy formulation and selection, implementation, evaluation, and revisiting choices—though actual policy development is rarely if ever so clean.[27] EBPM has a rich scholarly history,[28] buttressed by numerous models of how evidence informs policy,[29] and shows a keen awareness of how real-life decisions are likely to diverge sharply from the ideal. But the goal of using empirical evidence to improve

---

27.   CAIRNEY, *supra* note 23, at 17–18.

28*.   See, e.g.*, SMITH, *supra* note 18, at 5-6 (tracing EBPM from the 1960s onward); Linda Courtenay Botterill & Andrew Hindmoor, *Turtles All the Way Down: Bounded Rationality in an Evidence-Based Age*, 33 POL'Y RES. STUD. 376 (2011) (describing intellectual predecessors of EBPM in the 1950s).

29.   CAIRNEY, *supra* note 23, at 25; SMITH, *supra* note 18, at 10-20 (describing groups of theories to describe the process by which empirical evidence enters public policy-making).

policy outcomes continues to motivate improvements in both research methodology and the communication of research findings.[30]

The normative case for EBPM is predicated on several assumptions. First, regulators and legislators wish to promote the welfare of the people, and they act in good faith to achieve that goal. Efforts to improve the "effectiveness" of government in achieving its policy goals are therefore normatively desirable. Second, government resources are scarce, and some choices are mutually exclusive, such that regulators cannot simultaneously pursue all options. Third, empirical research can validly and reliably reveal facts about the effects of past policy responses, and this evidence is useful in predicting how current policy choices are likely to affect outcomes of interest. Fourth, it is possible to require and learn from contemporaneous evaluations of government choices. And finally, policy-makers are capable of identifying and learning from empirical evidence—or at least, capable of delegating those activities to others and acting on their recommendations.

When these assumptions hold, the case for EBPM is clear. Under conditions of limited resources (time, funds, capacity), it is both normatively desirable and efficient to allocate those resources to programs and policies most likely to promote citizens' well-being. Empirical evidence can thus improve the impact and efficiency of governmental choices.[31] The virtues of EBPM are that it is welfare-maximizing, that it can help allocate scarce resources efficiently, and that it is applicable regardless of the policy end in question. It thus has appeal regardless of party affiliation—whatever

---

30. One such effort is the continued improvement of methods for systematic reviews of empirical research, which aggregate findings across many studies and aim to communicate results in a policy-relevant way. Two international collaborations dedicated to the production and dissemination of systematic reviews are the Cochrane Collaboration and the Campbell Collaboration. CAMPBELL COLLABORATION ONLINE LIBRARIES, https://campbellcollaboration.org/library.html [https://perma.cc/NBT4-WTAT]; COCHRANE LIBRARY, https://www.cochranelibrary.com [https://perma.cc/BDA5-AYPL];.

31. Some criticize government efficiency wholesale. For example, Becker and Mulligan have advanced the argument that inefficient tax policy can be preferable because it forces citizens to notice and engage with government decisions that they may dislike. *See* Gary S. Becker & Casey B. Mulligan, *Deadweight Costs and the Size of Government*, 46 J.L. & ECON. 293 (2003) (noting that efficient taxes tend to promote the growth of government and suggesting that "an improvement in the efficiency of either taxes or spending would reduce political pressure for suppressing the growth of government").

the current policy goals may be, more effective pursuit of those goals is likely to be appealing to those in power.

Where the assumptions of EBPM do not hold—such as where government acts with invidious intent to disadvantage particular groups— greater efficiency can amplify the harm caused by legislative and regulatory action. Consider, for example, recent activity by the North Carolina legislature found to disenfranchise African American voters. Before designing new voter ID restrictions, legislators consulted data on the types of ID that black voters were less likely to have, such as licenses issued by the state DMV.[32] As the Fourth Circuit found, the legislature then relied on these data to tailor voter restrictions to disqualify the types of alternative ID used by black voters, "retain[ing] only the kinds of ID that white North Carolinians were more likely to possess."[33] The process of gerrymandering voting districts is similarly highly data-intensive and likewise an instance of EBPM in which legislators are motivated to collect data and to use the most robust available evidence to secure the outcomes they prefer. But where legislative and regulatory motives depart sharply from public well-being, improving the efficiency of government is no longer a good in itself. The approach of EBPM is thus an agnostic tool to promote efficiency, which is desirable only insofar as one supports the ends of governmental action.

B. The Limits of EBPM

Commentators have long hoped that EBPM will take the political rancor out of hard decisions, and recent calls for expanding EBPM have included bipartisan stories of compromise advanced by a mutual appreciation for

---

32. Christopher Ingraham, *The "Smoking Gun" Proving North Carolina Republicans Tried to Disenfranchise Black Voters*, WASH. POST (July 29, 2016), https://www.washingtonpost.com/news/wonk/wp/2016/07/29/the-smoking-gun-proving-north-carolina-republicans-tried-to-disenfranchise-black-voters/?utm_term=.6e20a69cdcc1 [https://perma.cc/ZE6V-7Z4E].

33. *Id.*; North Carolina State Conference of the NAACP v. McCrory, 831 F.3d 204 (4th Cir. 2016). Another illustration comes from legislative action to disenfranchise Native American voters in North Dakota by requiring street addresses rather than PO boxes, which took place immediately after Native American votes contributed to the election of Democratic Senator Heitkamp in 2012. The Supreme Court declined to bar the implementation of this law before the November 2018 election. Brakebill v. Jaeger, 586 U.S. ___ (2018).

empirical evidence.[34] Such compromises are unlikely, however, where there is deep-seated disagreement on what the goals of government should be. Moreover, even when policy-makers share a goal, rationalist views of policy-making are an imperfect match for real-world choices.[35] This Section considers four limitations that make EBPM a "bounded" rather than full-throated practice. There are many practical barriers to EBPM, including modifiable factors like collaborative relationships, access to research, costs of using research, and the clarity of research findings.[36] The following boundaries would persist, however, even after resolving these practical barriers, and they inevitably limit legislators and agencies' institutional capacity to practice EBPM thoroughly.

First, bounded rationality can impede even well-meaning attempts to use evidence effectively. Literature on irrational decision-making abounds,[37] and the decision makers in EBPM are as fallible as the rest of us,[38] particularly when making emotionally laden or moral decisions.[39] To name a few, these errors may include overestimating the likelihood of

---

34. *See* COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16; JIM NUSSLE & PETER ORSZAG, MONEYBALL FOR GOVERNMENT (2d ed. 2015); Haskins, *supra* note 23; John Bridgeland & Peter Orszag, *Can Government Play Moneyball?*, ATLANTIC (July/Aug. 2013), https://www.theatlantic.com/magazine/archive/2013/07/can-government-play-moneyball/309389 [https://perma.cc/6N3A-VJG3]; MONEYBALL FOR GOV'T, https://moneyballforgov.com [https://perma.cc/9Q4U-JRHF].

35. CAIRNEY, *supra* note 23, at 16 (noting the problems of a "linear" view of policy-making); Gamoran, *supra* note 10, at 185.

36. Kathryn Oliver et al., *A Systematic Review of Barriers to and Facilitators of the Use of Evidence by Policymakers*, 14 BMC HEALTH SERVS. 2, 7 (2014).

37. *See*, *e.g.*, DAN ARIELY, PREDICTABLY IRRATIONAL (2008); DANIEL KAHNEMAN, THINKING FAST AND SLOW (2011); CASS SUNSTEIN, SIMPLER: THE FUTURE OF GOVERNMENT (2013); RICHARD THALER & CASS SUNSTEIN, NUDGE (2009); Christine Jolls & Cass Sunstein, *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471 (1998).

38. *See* SUNSTEIN, *supra* note 37.

39. *See generally* CAIRNEY, *supra* note 23, at 26 (describing cognitive biases and the reliance on heuristics in emotional and moral reasoning); SUNSTEIN, *supra* note 37, at 69; Paul G. Lewis, Policy Thinking, Fast and Slow: A Social Intuitionist Perspective on Public Policy Processes (Aug. 19, 2013) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2300479 [https://perma.cc/K4XB-XGYF].

familiar or vividly imaginable events, regretting losses more acutely than we value gains, generalizing to social groups from individual examples, seeking out evidence that confirms our prior beliefs, changing our opinions depending on the framing of choices, updating beliefs to conform with others in our political party or social group, being stymied by ambiguity or complexity, interpreting emotions as information, believing that independent events are related, and believing that we will be luckier than others. Evidence from cultural cognition goes even further, demonstrating that we interpret the credibility of scientific research itself based on our cultural beliefs.[40] Decision-makers will also have predictable trouble sorting through empirical evidence alongside many other sources of information.[41]

Together, these biases often make it difficult or impossible to see research findings clearly, even with the best intentions to make evidence-based choices. These dynamics may exacerbate some of the pathologies that I will later describe, including technical breakdowns in the ability to use evidence, the distortion of evidence, and the reliance on nonrepresentative studies to make choices about larger programs (ratcheting). Sometimes the distortion of evidence is deliberate and for political ends. But sometimes even well-meaning regulators or legislators may misread, minimize, or elevate research findings because they confirm political priors or align with normative preferences. Moreover, even if policy debate were confined to empirical evidence, disputes about values and norms would infuse disputes about the selection and rigor of that evidence, with little loss of partisan rancor.

Beyond these predictable cognitive biases, the policy environment itself—with finite time, multiple demands for attention, many decision-makers, and high-stakes choices—affects both rational thinking and the capacity to receive and use nuanced information. Many decisions are made rapidly, under conditions of uncertainty and political pressure[42]—conditions where biases reign. Decision-makers' limited time and expertise

---

40. Dan M. Kahan et al., *Cultural Cognition of Scientific Consensus*, 14 J. RISK RES. 147 (2010).

41. Gamoran, *supra* note 10, at 185. These include past experience, personal values, political priorities, public opinion polls, anecdotes, lobbying efforts, and innumerable other sources of information. *See also* CAIRNEY, *supra* note 23, at 22 (noting that civil servants classify many types of information as "evidence," such as anecdotes, expert opinions, experiences from other governments, and public opinion).

42. CAIRNEY, *supra* note 23, at 22, 25.

also require the simplification of evidence,[43] in which relevant methodological caveats and nuanced distinctions are lost.[44] Modern theories of policy-making acknowledge these problems.[45] In 1959, for example, economist Charles Lindblom argued that policy-makers inevitably "muddle through" problems that require them to make choices, and that they often neglect better alternatives in the process.[46] Newer theories characterize policy development processes that are equally far afield from EBPM, even when they are amenable to *some* influence by empirical evidence.[47]

A second limitation of EBPM is the bounded ability to measure meaningful outcomes. Some values defy statistics. The problems of quantifying fundamental values—fairness, equality, autonomy, dignity— have long plagued efforts to develop evidence-based approaches to legislation and regulation. One of the steps the Obama administration took to advance EBPM was a January 2011 Executive Order, which directed agencies not only to engage in maximally accurate cost-benefit analysis, but to "consider (and discuss qualitatively) values that are difficult or impossible to quantify, including equity, human dignity, fairness, and distributive impacts."[48] Cass Sunstein, then-head of the Office of Information and Regulatory Affairs, describes both the ethical imperative and practical difficulty of addressing these concerns in the analysis of particular policies, including lifting the travel ban on HIV-positive individuals, proposing a rule to require rear-facing cameras in cars, strengthening EEOC regulations implementing the Americans with Disabilities Act, and issuing a potentially costly DOJ rule to reduce prison

---

43. *See* Botterill & Hindmoor, *supra* note 28.

44. *Id.*

45. CAIRNEY, *supra* note 23, at 25.

46. *See generally* Charles Lindblom, *The Science of Muddling Through*, 19 PUB. ADMIN. REV. 79, 81 (1959)(describing how idealized processes of policymaking are impossible under conditions of limited time, money, and information).

47. *See* SMITH, *supra* note 18, at 24–38 (describing modern theories that incorporate political economy among ruling elites, path dependency, rational choice theory, complexity theory, and theories of dramatic policy change).

48. Exec. Order No. 13563, 76 Fed. Reg. 3821, § 1(c) (Jan. 18, 2011). Regulatory impact analyses beginning in the Reagan administration also considered fairness as part of cost-benefit analysis. ANTHONY E. BOARDMAN ET AL., COST-BENEFIT ANALYSIS: CONCEPTS AND PRACTICE 20–21 (4th ed. 2017).

rape.[49] Even at its ideal, EBPM gives us little purchase on these choices—to the extent that EBPM purports to remove values-inflected debate and redirect it to a dispassionate focus on statistical evidence, the problems of "nonquantifiables"[50] would persist. This deficiency in evidence is perhaps always a form of technical breakdown in EBPM, but it may also contribute to evidence distortion (e.g., by providing external reasons to accept or reject research out of hand), ratcheting, or hesitancy to commit to long-term implementation of an important public policy (what I will call "terminal experimentation").

A third inherent boundary of EBPM is limits on scientific objectivity. Ideal EBPM is desirable in part because it promises a means of policy choice that is dispassionate, unaffected by political concerns. But this is an incomplete promise when scientific research itself embodies values and preferences—scientists, too, are boundedly rational, and boundedly objective.[51] Outright attacks on scientific knowledge as socially constructed came to a head in the 1990s "Science Wars" and subsequent debates over "junk science."[52] Although these debates have died down, they illuminated the ways in which scientific research often involves uncertainty and judgment—for instance, making inferences that generalize from past studies to current decisions[53] or judgments about generalizing findings to the nation as a whole. Researchers making such inferences will inevitably draw on values beyond objective scientific facts. Values, too, may drive the types of questions that scientists seek to answer,[54] the causal hypotheses they seek to test,[55] and the ways in which scientists frame their proposals and policy implications.[56] Where researchers do make assumptions,

---

49.   SUNSTEIN, *supra* note 37, at 166–168.

50.   *Id.* at 169.

51.   Regina Nuzzo, *How Scientists Fool Themselves – and How They Can Stop*, 526 NATURE 182 (2015).

52.   HEATHER DOUGLAS, SCIENCE, POLICY, AND THE VALUE-FREE IDEAL 5–13 (2009) (introducing the "Science Wars" and "junk science").

53.   *Id.* at 142.

54.   SMITH, *supra* note 18, at 177 (noting that research tends to be closely related to political and ideological outlook).

55.   Lawrence Gostin & Maxwell Gregg Bloche, *The Politics of Public Health: A Response to Epstein*, 46 PERSP. BIOLOGY & MED. S170 (2003) ("Inevitably, political, cultural, and moral influences shape the selection of targets.").

56.   SMITH, *supra* note 18, at 175.

scientific norms ask them to be explicit about their choices.[57] But it is difficult if not impossible to eradicate values entirely from evidence, and thus from EBPM. Drawing attention to these values can provide fodder for evidence distortion (e.g., shifting the focus from research findings to the researchers themselves) or terminal experimentation (e.g., citing uncertainty about research findings as a rationale for avoiding long-term policy commitments).

A fourth ineluctable limitation stems from the recognition that decisional processes are also a means of allocating power. Using empirical evidence in policy-making reinforces the authority of people who produce, interpret, and communicate that evidence,[58] which runs into problems of democratic legitimacy.[59] The processes of interest group politics are imperative for identifying distributional problems, threats to autonomy, or threats to dignity—all of which should pose real boundaries on EBPM, even if it means making decisions that are contrary to what empirical evaluations would counsel.[60] On this reading, a decision that deliberately ignores research evidence may be a technical breakdown, but it may also be a means of responding to a constituency with meaningful interests. Some of these interests may be nefarious. They may, for example, be rent-seeking or lobbying concerns that do not align with the public interest. But interest group politics are also important for identifying the needs of marginalized populations, and a version of EBPM that closes the door on political concerns is undesirable on democratic legitimacy grounds.

Idealized EBPM is unlikely even in the most congenial circumstances. But bounded EBPM can nonetheless make legislative and regulatory

---

57. DOUGLAS, *supra* note 52, at 155.

58. This calls to mind earlier claims of cost-benefit analysis as an arrogation of power by economists. Theodore M. Porter, *The Rise of Cost-Benefit Rationality as Solution to a Political Problem of Distrust* (unpublished manuscript), http://depts.washington.edu/econlaw/pdf/Porter.pdf [https://perma.cc/37MK-QBY3] (citing his book, which argues against such claims).

59. *See*, *e.g*., Gert Biesta, *Why "What Works" Won't Work: Evidence-Based Practice in the Democratic Deficit in Educational Research*, 57 EDUC. THEORY 1 (2007) (arguing that evidence-based practice in education policy limits opportunities for democratic participation in decision-making).

60. SMITH, *supra* note 18, at 3 (quoting Geoff Mulgan, *Government, Knowledge, and the Business of Policy-Making: The Potential and Limits of Evidence-Based Policy*, 1 EVIDENCE & POL'Y 215 (2005) ("[T]he people . . . have every right to ignore evidence.").

decisions *better*. For many choices, there is widespread agreement on what ends the legislature or agency should pursue, and there is a limited set of options that are both legal and feasible. Using the evidence that exists, where possible—despite limited rationality, measurement, objectivity of science, and authority—can help to make those choices better than before.

C. Bounded EBPM Rising: From the Census to Moneyball

Accepting the boundaries above, "a realistic goal is to have evidence present for consideration rather than assuming it will dominate the decision process."[61] By those lights, bounded EBPM is gaining attention. The following brief history of EBPM will discuss its advancement in federal agency regulation and, more recently, in Congressional efforts to improve the collection and aggregation of statistical data to inform legislative action.

1. Origins and Cost-Benefit Analysis

Histories of federal evidence-based policy in the U.S. differ in their starting points. Some would date the interaction between science and policy back to the framers' constitutional requirement that a census of the population be conducted every 10 years,[62] along with James Madison's argument that census results would be useful for congressional action.[63] Others trace the US origins of EBPM to the creation of the National Academy of Sciences to provide scientific advice during the Civil War,[64] or the

---

61.   Gamoran, *supra* note 10, at 185; *see also* CAIRNEY, *supra* note 23, at 75 ("[W]e need to move away from the idea that policy is made from the top down; that the best evidence, derived from "gold standard" methods, feeds directly into the top, and its insights are used in a straightforward implementation process at the bottom. If policy is a messier process involving multiple actors and levels of government, and it seems to 'emerge' from the interaction between actors and local levels, we need some way to inject evidence into *that* process.").

62.   U.S. CONST. art. I, § 2 ("The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.").

63.   COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16, at 12.

64.   DOUGLAS, *supra* note 52, at 25.

National Research Council as an arm of the NAS during World War I.[65] Most accounts of the modern EBPM movement, however, pick up during and after World War II, with landmark actions such as President Roosevelt's 1939 charge to the Bureau of the Budget to "promote the improvement . . . of Federal and other statistical services,"[66] the Food and Drug Administration's new requirements for randomized, placebo-controlled drug trials in 1962,[67] and the establishment of new administrative infrastructure for scientific advising[68] and statistical data collection.[69] The 1960s and 1970s saw further developments including the provisions for mandatory evaluations of governmental programming in the Departments of Education, Health and Human Services, and Labor,[70] the creation of advisory bodies such as Assistant Secretary for Planning and Evaluation in HHS,[71] the academic recognition and expansion of randomized trials for studying social programming,[72] and President Nixon's reshaping of the Bureau of the Budget as the Office of Management and Budget (OMB) in

---

65. *Id.* at 26; *History*, NAT'L ACAD. SCI., http://www.nasonline.org/about-nas/history/archives/milestones-in-NAS-history/organization-of-the-nrc.html [https://perma.cc/L6HG-EDEU].

66. COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16, at 12–13.

67. Jon Baron, *A Brief History of Evidence-Based Policy*, 678 ANNALS AAPSS 40, 42 (2018).

68. These included National Defense Research Committee and the Office of Scientific Research and Development. DOUGLAS, *supra* note 52, at 30–32. This time also saw the establishment of the Council of Economic Advisers in 1946 and the Office of Science and Technology Policy in 1976. COUNCIL ECON. ADVISERS, https://www.whitehouse.gov/cea [https://perma.cc/KYU3-T4P4]; OFFICE SCI. TECH. POLICY, https://www.whitehouse.gov/ostp [https://perma.cc/F72Q-VUSH].

69. *E.g.*, Federal Records Act of 1942, Pub. L. No. 81-754, 64. Stat. 578 (1950) (codified at 44 U.S.C. § 3101 et seq.) (allocating authority for review of data collection in government).

70. Rebecca A. Maynard, *The Role of Federal Agencies in Creating and Administering Evidence-based Policies*, 678 ANNALS AAPSS 134, 134 (2018); Larry L. Orr, *The Role of Evaluation in Building Evidence-Base Policy*, 678 ANNALS AAPSS 51, 52–54 (2018).

71. OFFICE ASSISTANT SEC'Y FOR PLANNING & EVALUATION, https://aspe.hhs.gov [https://perma.cc/W7MN-X9PS].

72. Baron, *supra* note 67, at 43-44; Head, *supra* note 23, at 78.

1970[73]—the office that would later "quarterback" evidence-based practice in federal administration.[74]

Efforts to integrate empirical evidence in federal policy, however, owe most to the 1980s-era entrenchment of cost-benefit analysis (CBA) and regulatory impact statements throughout the executive branch. Although the federal government had tested CBA in the U.S. Army Corps of Engineers in the 1930s,[75] the practice was not mandated until President Reagan's executive order in 1981, which required a regulatory impact analysis for all "major rules."[76] Analyses had to consider potential benefits,[77] potential costs, net benefits, alternative approaches, and legal constraints.[78] President Clinton followed with a parallel executive order in 1993,[79] and CBA has since become a routine part of agency decision-making.[80] CBA is not in itself evidence-based practice—that is, regulatory impact analyses are not necessarily *used* to drive policy choices. But both presidential administrations also took steps to require EBPM where possible. Reagan tasked the newly created OIRA with the oversight of federal regulation, including the requirement that regulatory action should not proceed "unless the potential benefits . . . outweigh the potential cost to society."[81] Surely regulators had always believed that their actions were likelier than not to benefit society. But with CBA, there was a structure requiring them to examine available evidence before reaching that conclusion.

The uses of rigorous study designs to evaluate policy and programmatic choices also expanded greatly in the 1970s and 1980s, with landmark

---

73. Exec. Order No. 11541, 35 Fed. Reg. 10737 (July 1, 1970).

74. Kathy Stack, *The Office of Management and Budget: the Quarterback of Evidence-Based Policy the Federal Government*, 678 ANNALS AAPSS 112 (2018). An important step in this development was the Paperwork Reduction Act of 1980, which accorded OMB authority to develop and coordinate statistical activities across the federal government. COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16, at 14.

75. BOARDMAN ET AL., *supra* note 48, at 20-21.

76. *Id.* at 21 (citing Executive Order 12291).

77. Agencies were directed to consider not only financial benefits, but also "any beneficial effects that cannot be quantified in monetary terms." *Id*.

78. Exec. Order No. 12291, 46 Fed. Reg. 13193 (Feb. 17, 1981).

79. Exec. Order No. 12866, 58 Fed. Reg. 51735 (Oct. 4, 1993).

80. BOARDMAN ET AL., *supra* note 48, at 21; *see* Matthew D. Adler & Eric A. Posner, *Rethinking Cost-Benefit Analysis*, 109 YALE L.J. 165, 167 (1999).

81. Exec. Order No. 12291, *supra* note 78; *see* SUNSTEIN, *supra* note 37, at 2.

government-sponsored studies such as the RAND Health Insurance Experiment funded by HHS (then known as the Department of Health, Education, and Welfare),[82] a set of DOL-led studies on bonuses for job-seekers who obtained employment before the conclusion of their unemployment benefits,[83] the completion of evaluations of negative income tax (income maintenance programs),[84] and the welfare-to-work experiments of the 1980s.[85]

2. The Modern Era

Outside the U.S., EBPM received a public boost in 1997 when the New Labour party swept the UK polls, bringing its "what works" manifesto to the parliamentary majority and prime minister's seat.[86] Stateside, "evidence-based practice" and "EBPM" gained recognition throughout the 1990s, as did efforts to evaluate government performance.

One such effort was the Government Performance and Results Act of 1993 (the GPRA, later updated in 2010), by which Congress directed agencies to establish goals and report on their performance.[87] Federal efforts identifying effective programs emerged in the 1990s and 2000s, including the What Works Clearinghouse in the Education Department,[88] the National Commission on Families in HHS,[89] and the listings of Evidence-Based Interventions at the Centers for Disease Control and Prevention.[90] Agencies established new research and advisory units, such as the Institute

---

82. *RAND's Health Insurance Experiment (HIE)*, RAND CORP., https://www.rand.org/health/projects/hie.html [https://perma.cc/7LVX-TDHJ].

83. GREENBERG ET AL., SOCIAL EXPERIMENTATION AND PUBLIC POLICYMAKING 166-169 (2003).

84. *Id.* at 111–164 (describing a series of studies that took place during 1968–1980).

85. *Id.* at 211–269 (noting that the findings were largely neglected in PRWORA).

86. SMITH, *supra* note 18, at 11–12; Botterill & Hindmoor, *supra* note 28, at 369.

87. COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16, at 14–15.

88. Maynard, *supra* note 70, at 137.

89. *Id.*

90. *Compendium of Evidence-Based Interventions and Best Practices for HIV Prevention*, CTR. DISEASE CONTROL & PREVENTION, https://www.cdc.gov/hiv/research/interventionresearch/compendium/index.html [https://perma.cc/B9YW-KZQN].

of Education Sciences in the Education Department.[91] The 1980s and 1990s also saw an increase in randomized trials to evaluate experimental welfare programs, inspired in part by a series of welfare-to-work trials carried out by Manpower Demonstration Research Corporation in the early 1980s.[92] In the Family Support Act of 1988, Congress required randomized evaluations of experimental AFDC welfare programming carried out by states.[93] Notably, these trials embedded random assignment of welfare beneficiaries into ordinary operations carried out by state agencies, rather than evaluating bespoke programs intended for scientific study.[94] The 1990s also saw randomized trials of other policy interventions aimed at poverty reduction, such as the Moving to Opportunity housing study sponsored by HUD and approved by Congress in 1992.[95]

Agency efforts to integrate empirical evidence into decision-making expanded further under President George W. Bush. The President's Management Agenda released by OMB in 2001 emphasized "results-oriented" government, culminating in the Program Assessment Rating Tool (PART) announced in 2002.[96] Under the PART initiative, OMB directed agencies to evaluate all programming once per five years using a survey instrument, and further, noted that randomized controlled trials were preferred wherever feasible.[97] Over five years, OMB used the PART survey instrument to assess federal programs constituting ninety-eight percent of

---

91.   INST. EDUC. SCI., https://ies.ed.gov [https://perma.cc/UGZ5-4PNN].

92.   Baron, *supra* note 67, at 43; Carol Harvey et al., *Evaluating Welfare Reform Waivers Under Section 1115*, 14 J. ECON. PERSP. 165, 175 (2000).

93.   Family Support Act of 1988, Pub. L. No. 100-485, § 487(BXi)(iii), 102 Stat. 2343, 2380 (Oct. 13, 1988) ("A demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of participants in the program . . . ."); Baron, *supra* note 67, at 44.

94.   Baron, *supra* note 67, at 44; Harvey et al., *supra* note 92, at 172.

95.   *A Summary Overview of Moving to Opportunity: A Random Assignment Housing Mobility Study in Five U.S. Cities*, NAT'L BUREAU OF ECON. RESEARCH, http://www.nber.org/mtopublic/MTO%20Overview%20Summary.pdf (last visited Dec. 21, 2018) [https://perma.cc/BX2H-MHY8].

96.   RON HASKINS & JON BARON, BUILDING THE CONNECTION BETWEEN POLICY AND EVIDENCE: THE OBAMA EVIDENCE-BASED INITIATIVES 8 (2011); Orr, *supra* note 70, at 55.

97.   Orr, *supra* note 70, at 55.

the federal budget.[98] The PART initiative was not uncontroversial; although PART ratings were not exclusively determinative of funding,[99] many Democrats and agency staff interpreted the assessment tool as laying the groundwork for Republican budget cuts to social programming.[100] Subsequent analyses have found that PART scores were significantly correlated with Bush's proposed budget, but that the scores did not appear to influence Congressional appropriations.[101] The tool itself was discontinued by President Obama.[102] But in its place, the Obama administration instituted a new set of agency evaluation practices, including with OMB instructions that agencies should support budget requests with randomized controlled trials where available.[103]

Several detailed accounts have outlined the robust expansion of EBPM under the Obama presidency, the administration most aligned with the modern evidence-based practice movement.[104] As Ron Haskins and Jon Baron wrote in 2011, "The Obama administration . . . created a sweeping new opportunity for rigorous evidence to influence policy. No president . . . [has] ever been so intent on using evidence to shape decisions about the funding of social programs."[105] During the Obama years, OMB and OIRA took

---

98. Donald P. Moynihan, *Advancing the Empirical Study of Performance Management: What We Learned from the Program Assessment Rating Tool* 6 (La Follette Sch. Working Paper No. 2013-003), http://www.lafollette.wisc.edu/images/publications/workingpapers/moyni han2013-003.pdf [https://perma.cc/MMB4-G67Z].

99. *Id.* at 9.

100. *See, e.g.*, Stack, *supra* note 74, at 115-120.

101. Moynihan, *supra* note 98, at 9.

102. Patrick Lester, *Managing Toward Evidence: State-Level Evidence-Based Policymaking and the Results First Initiative*, 678 ANNALS AAPSS, 93, 97 (2018); *see also* Moynihan, *supra* note 98, at 6 ("The Obama administration characterized the tool as ineffective at generating true performance information.").

103. Orr, *supra* note 70, at 55.

104. *See, e.g.*, HASKINS & BARON, *supra* note 96; RON HASKINS & GREG MARGOLIS, SHOW ME THE EVIDENCE: OBAMA'S FIGHT FOR RIGOR AND RESULTS IN SOCIAL POLICY (2014); SUNSTEIN, *supra* note 37; William J. Congdon & Maya Shankar, *The Role of Behavioral Economics in Evidence-Based Policymaking*, 678 ANNALS AAPSS 81 (2018); Stack, *supra* note 74.

105. Ron Haskins & Jon Baron, *Building the Connection between Policy and Evidence: The Obama Evidence-Based Initiatives* 6-7 (2011),

central roles in applying scientific evidence and prioritizing agency performance, led respectively by Peter Orszag and Sunstein.[106] In 2011, Obama's executive order "Improving Regulation and Regulatory Review" emphasized that policy must be "based on the best available science," and "must measure, and seek to improve, the actual results of regulatory experiments."[107] The order mandated retrospective review of existing rules[108] and directed agencies to "ensure the objectivity of any scientific and technological information and processes used to support the agency's regulatory actions."[109] Another executive order directed agencies to incorporate and test their applications of insights from the field of behavioral economics, and established the Social and Behavioral Sciences Team to summarize agency progress toward this goal.[110]

Congress also developed a greater interest in EBPM during these years, establishing a large number of pilot programs and demonstration projects with evaluation mandates attached, such as the No Child Left Behind Act of 2001 (an early example that prioritized randomized trials[111]); the Affordable Care Act of 2010; the Workforce Innovation and Opportunity Act of 2014; the Agricultural Act of 2014; the Bipartisan Budget Act of 2015; and the Every Student Succeeds Act in 2015.[112] In 2016, Congress authorized the Commission on Evidence-Based Policymaking to explore opportunities for

---

http://coalition4evidence.org/wp-content/uploads/2011/09/Haskins-Baron-paper-on-fed-evid-based-initiatives-2011.pdf [https://perma.cc/SX49-KLAH].

106. *See* SUNSTEIN, *supra* note 37 (describing the role of OIRA); Stack, *supra* note 74 (describing the leadership of OMB in advancing EBPM in the Obama administration).

107. Exec. Order No. 13563, 76 Fed. Reg. 3821 (Jan. 18, 2011).

108. *Id.* at 3822.

109. *Id.*

110. Exec. Order No. 13707, 3 C.F.R. 13707 (Sept. 15, 2015). This mirrors the Behavioral Insights Team ("Nudge Unit") in the UK, which still operates with UK government support. *See also* Congdon & Shankar, *supra* note 104 (describing the work of the SBST).

111. Gamoran, *supra* note 10; Debra Viadero, *U.S. Position on Research Seen in Flux*, EDUC. WEEK (Mar. 4, 2008), https://www.edweek.org/ew/articles/2008/03/05/26research_ep.h27.html [https://perma.cc/2JRY-M4EU].

112. COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16, at 15.

using and improving EBPM in practice.[113] Soon after the Commission's report, the House passed the Foundations for Evidence-Based Policymaking Act, seeking to improve "evidence-building activities" coordinated by OMB;[114] a Senate version soon followed,[115] and the bill became law early in 2019.[116] The 2017 Congress considered several other proposed bills intended to improve the generation and uses of data, including the Legislative Performance Review Act—a bill that would require both houses of Congress to review GPRA data in budget appropriations decisions.[117] State legislators[118] and municipal governments[119] have also incorporated EBPM aspirations and practices in tandem with progress at the federal level.

It is early to appraise how evidence-based practice will fare overall under the Trump administration, but the administration demonstrably has a different emphasis.[120] A hallmark of the rise of populist government

---

113. Evidence-Based Policymaking Commission Act of 2016, Pub. L. No. 114-140, 130 Stat. 317 (2016); COMM'N ON EVIDENCE-BASED POLICYMAKING, *supra* note 16.

114. Foundations for Evidence-Based Policymaking Act of 2017, H.R. 4174, 115th Cong. (2017).

115. Foundations for Evidence-Based Policymaking Act of 2017, S. 2046, 115th Cong. (2017).

116. Foundations for Evidence-Based Policymaking Act of 2017, Pub. L. No. 115-435, 132 Stat. 5529 (2019) (directing OMB to coordinate an "evidence-building" plan each year to address "policy-relevant" questions from federal agencies; establishing a government-wide Interagency Council on Evaluation Policy; directing OMB to create an Advisory Committee on Data for Evidence Building; requiring each agency to name a Chief Evaluation Officer and a Chief Data Officer; and requiring each agency to develop a data inventory as well as to publish open data in machine-readable format).

117. 1 EVIDENCE USE IN CONGRESS, BIPARTISAN POLICY CTR. 12 (Mar. 2018), https://bipartisanpolicy.org/wp-content/uploads/2019/03/BPC-Evidence-Use-in-Congress.pdf [https://perma.cc/B38T-QA5L].

118. Jeremy L. Hall & Edward T. Jennings Jr., *Assessing the Use and Weight of Information and Evidence in U.S. State Policy Decisions*, 29 POL'Y & SOC'Y 137, 137 (2010).

119. WHAT WORKS CITIES, https://whatworkscities.bloomberg.org [https://perma.cc/XQ7F-DLWK].

120. In the words of one commentator: "President Obama's regulatory Czar Cass Sunstein declared that we had moved beyond pro- and anti-regulation polarization into rule by experts, where 'state-of-the-art techniques for anticipating, cataloguing, and monetizing the consequences of regulation' answer seemingly every major question. I think it is fair to say that this is

approaches in both the US and the UK has been an aversion to technocratic expertise.[121] As UK Brexit advocate Michael Gove summed it up, "people in this country have had enough of experts."[122] Salient Obama-era EBPM programming has been discontinued, such as the behavioral economics initiative in OMB.[123] Trump's early executive order on regulation required the repeal of two administrative regulations for every new one passed, a strategy far from an ideal EBPM playbook.[124]

Scientists have overwhelmingly expressed concern about neglect of empirical evidence in the administration.[125] Indeed, several social programming decisions to date have made selective use of evidence by "cherrypicking studies" and emphasizing uncertainty to de-fund programs.[126] The administration has expanded funding for approaches that have been rejected in systematic reviews of scientific literature, such as

exactly the kind of mindset that Trump ran against. And it is exactly that kind of mindset that the Trump Administration, in its early days, is repudiating." Andrew M. Grossman, *An Administration Takes Sides* YALE J. REG.: NOTICE AND COMMENT, (May 25, 2017), http://yalejreg.com/nc/an-administration-takes-sides-by-andrew-m-grossman [https://perma.cc/7MEU-5PGH].

121. *See, e.g.*, William Davies, *Why We Stopped Trusting Elites*, GUARDIAN (Nov. 29, 2018), https://www.theguardian.com/news/2018/nov/29/why-we-stopped-trusting-elites-the-new-populism [https://perma.cc/4TBW-34HT] (arguing that trust in public institutions has declined in the UK); Cary Funk et al., *Trust and Mistrust in Americans' Views of Scientific Experts*, PEW RES. CTR. (2019), https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts/ [https://perma.cc/D9RX-4U2L] (finding that 73% of Democrats but only 43% of Republicans believe that scientists should be active in "scientific policy debates").

122. Steven Erlanger, *No Charisma, No Glamour: Michael Gove Makes His Case to Lead Britain*, N.Y. TIMES (July 2, 2016), https://www.nytimes.com/2016/07/02/world/europe/brexit-michael-gove-conservative-party-britain.html [https://perma.cc/LK8Q-9JY3].

123. Haskins, *supra* note 23, at 19.

124. Exec. Order No. 13771, 82 Fed. Reg. 9339 (Feb. 3, 2017).

125. *See*, *e.g.*, MARCH FOR SCIENCE, https://www.marchforscience.com [https://perma.cc/RPX7-5RUW].

126. Robert Gordon & Ron Haskins, *The Trump Administration's Misleading Embrace of 'Evidence'*, POLITICO (Mar. 31, 2017, 5:35 AM), https://www.politico.com/agenda/story/2017/03/the-trump-administrations-misleading-embrace-of-evidence-000385 [https://perma.cc/2LL5-RN7E].

funding for abstinence-based sex education.[127] The phrases "evidence-based" and "science-based" were standouts on an infamous 2017 list of politically contentious phrases that CDC officials were instructed to avoid, perhaps as a means of "self-preservation."[128] Scientific advisory boards have met less frequently under the Trump administration than in prior years,[129] with more cancelled meetings and fewer scientific expert members than in prior administrations. Notable committee closures have included the Department of the Interior's Advisory Committee on Climate Change and Natural Resource Science,[130] the FDA's Food Advisory Committee,[131] the EPA's Particulate Matter Review Panel,[132] the Advisory Committee for the Sustained National Climate Assessment,[133] and several panels that

127. *See* John S. Santelli et al., *Abstinence-Only-Until-Marriage: An Updated Review of U.S. Policies and Programs and Their Impact*, 61 J. ADOLESCENT HEALTH 273 (2017); Kristen Underhill et al., *Sexual Abstinence Only Programmes to Prevent HIV Infection in High Income Countries: Systematic Review*, BMJ, July 26, 2007, at 1; Cheryl Wetzstein, *Sex-Education Funds to Be Renewed*, WASH. TIMES (Apr. 15, 2015), https://www.washingtontimes.com/news/2015/apr/15/sex-education-funds-to-be-renewed [https://perma.cc/MM4M-VS9V].

128. Instead, CDC personnel have reportedly been instructed to say, "CDC bases its recommendations on science in consideration with community standards and wishes"—a movement in terminology that may well capture a more realistic view of policy development. Brakkton Booker, *Trump Administration Reportedly Instructed CDC on Its Own Version of 7 Dirty Words*, NPR (Dec. 16, 2017), https://www.npr.org/2017/12/16/571329234/trump-administration-reportedly-instructs-cdc-on-its-own-version-of-7-dirty-word [https://perma.cc/GU27-S22C].

129. UNION OF CONCERNED SCIENTISTS, ABANDONING SCIENCE ADVICE: ONE YEAR IN, THE TRUMP ADMINISTRATION IS SIDELINING SCIENCE ADVISORY COMMITTEES (2018), https://www.ucsusa.org/sites/default/files/attach/2018/01/abandoning-science-advice-full-report.pdf [https://perma.cc/H7PN-BMD7].

130. *Id.*

131. *Id.*

132. Lisa Friedman, *E.P.A. to Disband a Key Scientific Review Panel on Air Pollution*, N.Y. TIMES (Oct. 11, 2018), https://www.nytimes.com/2018/10/11/climate/epa-disbands-pollution-science-panel.html [https://perma.cc/4RL4-AHTW].

133. Juliet Eilperin, *The Trump Administration Just Disbanded a Federal Advisory Committee on Climate Change,* WASH. POST (Aug. 20, 2017), https://www.washingtonpost.com/news/energy-environment/wp/

advised OSHA.[134] Other scientific panels face new limitations, such as the Department of Energy's Defense Nuclear Facilities Safety Board, which has newly restricted abilities to access information and receive responses from DOE contractors.[135]

A few agency decisions and executive orders, however, have also made some moves towards promoting randomized trials. The Executive Order Expanding Apprenticeships in America requests that OMB oversee a comprehensive evaluation of all "job training programs" administered by federal agencies, noting, "When feasible, these evaluations shall be conducted by third party evaluators using the most rigorous methods appropriate and feasible for the program, with preference given to multi-site randomized controlled trials."[136] The Trump HHS has released new recommendations to guide states in their evaluations of experimental Medicaid programming, including enforcing prior recommendations to ensure the independence of evaluators,[137] and has been supportive of statewide randomized trials for evaluating Medicaid waivers.[138] (The Trump HHS has also cancelled long-term data collection for one such

2017/08/20/the-trump-administration-just-disbanded-a-federal-advisory-committee-on-climate-change [https://perma.cc/EAD2-RE2V].

134. Rebecca Moss, *Trump's Labor Department Eviscerates Workplace Safety Panels,* SANTA FE NEW MEXICAN (Mar. 29, 2018), https://www.propublica.org/article/trumps-labor-department-eviscerates-workplace-safety-panels [https://perma.cc/UVJ5-XYWJ].

135. U.S. DEP'T OF ENERGY, DOA 0140.1, INTERFACE WITH THE DEFENSE NUCLEAR FACILITIES SAFETY GUARD (2018), https://www.directives.doe.gov/directives-documents/100-series/0140.1-BOrder/@@images/file [https://perma.cc/GL2B-FKE7]; Rebecca Moss, *Trump Administration Neuters Safety Board*, PROPUBLICA (Jul. 22, 2018), https://www.propublica.org/article/nuclear-safety-board-information-access-trump-administration [https://perma.cc/LLK9-F6EC].

136. Exec. Order No. 13801, 82 Fed. Reg. 28229 (June 15, 2017).

137. *Section 1115 Demonstrations Developing the Evaluation Design Advisory Recommendations*, CTR. FOR MEDICAID & MEDICARE, https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/developing-the-evaluation-design.pdf [https://perma.cc/V6NF-8R8T].

138. *Evaluation of the Health and Economic Consequences of Kentucky's Section 1115 Demonstration Waiver*, CLINICALTRIALS.GOV, https://clinicaltrials.gov/ct2/show/NCT03602456 [https://perma.cc/2YCJ-L3DY].

waiver, however.[139]) If Trump's order requiring the repeal of two regulations for every one issued[140] provokes a retrospective review of regulation effectiveness, it may be a catalyst for EBPM, but this order contained no criteria for deciding which regulations are axed.[141]

3. "Moneyball for Government"

In place of "EBPM," the catchier term "regulatory moneyball" is gaining ground in popular discourse.[142] The phrase derives from Michael Lewis's account of Billy Beane, general manager for the Oakland A's, who brought the team to glory by rejecting traditional scouting wisdom in favor of statistical evidence of players' skill. "Moneyball for government" promotes a similar shift toward research evidence as a means of decision-making, including "build[ing] evidence about the practices, policies, and programs that will achieve the most effective and efficient results," and investing public money accordingly.[143] This is unmistakably a rebranding of EBPM. Just as moneyball transformed the Oakland A's, proponents argue that "data, evidence, and evaluation [can] similarly revolutionize America's government . . . by investing in what works, by testing it and retesting it, and holding ourselves to a higher standard."[144] Other efforts abound: the What Works Cities initiative publicly evaluates and praises municipal governments that use data to improve local decisions;[145] the Pew-

---

139. Virgil Dickson, *CMS Scales Back Evaluation Plans for Indiana Medicaid Waiver*, MODERN HEALTHCARE (Apr. 9, 2018), https://www.modernhealthcare.com/article/20180409/NEWS/180409915 [https://perma.cc/2GLG-822X].

140. Exec. Order. No. 13,771, 82 Fed. Reg. 9339 (January 30, 2017).

141. Connor Raso, *Assessing Regulatory Retrospective Review Under the Obama Administration*, BROOKINGS, (June 15, 2017), https://www.brookings.edu/ research/assessing-regulatory-retrospective-review-under-the-obama- administration/ [https://perma.cc/K967-KLLH].

142. *See*, *e.g.*, MONEYBALL FOR GOV'T, *supra* note 34 (naming "all-star" legislators, regulators, and academics committed to advancing the uses of empirical evidence in policy); NUSSLE & ORSZAG, *supra* note 34; SUNSTEIN, *supra* note 37, at ch. 10; Bridgeland & Orszag, *supra* note 34.

143. Nussle & Orszag, *Let's Play Moneyball*, *in* NUSSLE & ORSZAG, *supra* note 34, at 6.

144. *Id*. at 3–4.

145. *About*, WHAT WORKS CITIES, BLOOMBERG, https://whatworkscities .bloomberg.org/about [https://perma.cc/9Q44-PMFQ].

MacArthur Results First initiative supports states seeking to use EBPM;[146] and organizations such as J-PAL,[147] Results for America,[148] the Coalition for Evidence-Based Policy,[149] the open-data efforts of the Sunlight Foundation,[150] and the continued Behavioral Insights Team in the UK[151] are visible nongovernmental efforts to promote EBPM in practice.

II. BROKEN EXPERIMENTATION, SHAM EBPM

The positivist ideal of data-driven government is elusive and often undesirable, but bounded EBPM is both feasible and salutary where policy ends are benign. EBPM has also been supported by growing congressional attention and, at least in foregoing administrations, agency efforts to implement EBPM with evaluation. But despite these intentions, current practice of EBPM in legislation and regulation is hit-or-miss.

This Part will turn an original focus to ways in which efforts that resemble EBPM on their face can go wrong. Indeed, I argue that sometimes these actions go so badly awry that they are not simply incomplete, but are, instead, *sham* versions of evidence-based practice. Breakdowns affect the use of evidence in primary decision-making, as well as the generation of evidence to evaluate those choices. I will begin with the latter category of broken experimentation, highlighting some of the design and implementation flaws that undermine government-led or government-sponsored evaluations of policy choices. I then argue that there are distinct pathologies in the link between evidence and decisions (sham EBPM), ranging from technical breakdowns, to willful neglect, to more disingenuous uses of "experimentation" and "evidence-based" nomenclature to achieve values-driven ends.

---

146. *Pew-MacArthur Results First Initiative*, PEW TRUST, http://www.pewtrusts .org/en/projects/pew-macarthur-results-first-initiative [https://perma.cc/U2G4-QML3].

147. J-PAL, https://www.povertyactionlab.org [https://perma.cc/9FYJ-24F8].

148. RESULTS FOR AM., http://results4america.org [https://perma.cc/GG7S-2DJY].

149. HASKINS & BARON, *supra* note 96.

150. SUNLIGHT FOUND., https://sunlightfoundation.com [https://perma.cc/557L-HACE].

151. BEHAVIORAL INSIGHTS TEAM, https://www.behaviouralinsights.co.uk [https://perma.cc/MQ8J-QS8S].

A. Broken Experimentation

The pathologies of evidence production are many, and no research field is without serious concerns about research methods and the validity and reliability of research findings. The evaluation of policy choices is no exception. In government-led and government-sponsored evaluation of policy choices, robust evaluation has tended to be the exception rather than the rule, which has been a stumbling block for efforts to learn from policy interventions. Failures to embed evaluations with policy choices, along with methodological flaws present in evaluations that are done, mean that evidence necessary for well-functioning EBPM is absent or, more worryingly, misleading.

At the outset, many evaluations simply do not occur. They are time-consuming and expensive, and there is an understandable emphasis on using government funds to accomplish the goals of a program, rather than to evaluate whether the program works as intended. As Orszag has noted, evaluations "are typically an afterthought when programs are designed—and once programs have been in place for a while, evaluating them rigorously becomes difficult from a political economy perspective."[152] A 2013 GAO survey of agency personnel in charge of programming found widespread support for using evaluation results, but only thirty-seven percent of personnel reported that *any* program they oversaw had been evaluated in the past five years.[153] The absence of evaluation hampers efforts to revisit and update policy choices over time, which is an important

---

152. Pater Orszag, *Building Rigorous Evidence to Drive Policy,* OFFICE OF MGMT. & BUDGET (June 8, 2009), https://obamawhitehouse.archives.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy [https://perma.cc/P7Y9-MBLN].

153. Of the group who reported past evaluations of agency programs, however, a majority found those evaluations "moderately" or "greatly" useful for implementing or planning program changes, assessing effectiveness, understanding and explaining program impacts, sharing information with others, developing program goals, allocating resources, streamlining programs, and supporting budget requests. Notably, concerns about evidence relevancy and credibility, evidentiary disagreements, lack of regulators' commitment to research, and ability to use evidence were consistently ranked as "small" or "none." U.S. GOV'T ACCOUNTABILITY OFFICE, GAO 13-570, PROGRAM EVALUATION: STRATEGIES TO FACILITATE AGENCIES' USE OF EVALUATION IN PROGRAM MANAGEMENT AND POLICY MAKING 15 (2013), https://www.gao.gov/assets/660/655518.pdf [https://perma.cc/5S3Z-QH85].

capacity given the limitations of using past evidence to predict future policy effects.

When government-led policy evaluations do occur, they are often limited to performance evaluations (rather than impact evaluations) and lack control groups. Although a one-group, pre-post study is useful for some questions, such as whether it is administratively feasible to deliver a program, it does very little to assess the impact of a program on outcomes of interest. And where evaluations use control groups, many miss the opportunity to use random assignment to enable an assessment of program effects. Randomized trials derive their methodological strength from the fact that individual participants are assigned to the treatment vs. control group purely on the basis of chance, rather than on the basis of any characteristic that might lead to systematic differences among the groups.[154] Differences between the groups after the policy is implemented, therefore, can be conclusively linked to the policy itself, rather than any other known or unknown confounder.

Despite the advantages of randomization, there is often resistance to using this methodological tool. This reluctance often reflects strong normative commitments in favor of (or against) a policy choice—namely, the belief that it would be inequitable to withhold a presumed benefit from one group (or to inflict a presumed harm on one group) on the basis of chance.[155] Legal concerns about randomization have generally failed in judicial challenges to randomized studies.[156] Randomization may indeed be ethically questionable if there is not genuine uncertainty about whether a new or existing policy is more effective. Here, policy-makers' and researchers' normative commitments—which are part of the bounded objectivity of science described above—may limit the choices of research design.

But despite past resistance to the use of randomization to evaluate policy impacts,[157] randomized trials are gaining new attention,[158] as are methods that approximate but do not perfectly comply with randomization.

---

154. Michael Abramowicz et al., *Randomizing Law*, 159 U. Pa. L. Rev. 929 (2011); Gordon H. Guyatt et al., *Users' Guide to the Medical Literature XXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care*, 284 JAMA 1290 (2000).

155. Colleen Chien, *Rigorous Policy Pilots*, 104 Iowa L. Rev 2313, at 2318 (2019).

156*. Id.*

157. Abramowicz et al., *supra* note 154.

158. Sunstein, *supra* note 37, at 189.

These include staged rollouts of programs (which allow a temporary control group to get the program after a delay) or implementation of policies depending on cutoff points (e.g., benefits on the basis of a federal poverty line cutoff).[159] Randomization has its scientific critics,[160] but where trials are large, attentive to subgroups and outcomes of interest, and reported appropriately, they are the most rigorous means of assessing the causal impact of interventions.[161]

Randomization is not the end of the matter. Even evaluations that use random assignment will fail to measure program impacts if they are statistically underpowered —namely, if there are too few people enrolled in the study to detect differences between people who receive the intervention and people who do not.[162] Small evaluations are cheaper and more easily controlled, and they exempt fewer people from policies that regulators or the legislature expect to be helpful. But where they lack statistical power to identify the impacts of the program, they are a waste of evaluation resources and will be biased toward showing a lack of program effectiveness.

Another pitfall for evaluations of governmental programs is a failure to assess outcomes of importance, including outcomes of importance for the individuals who experience the policy. This verges on a democratic legitimacy issue. Ideally, the design of a policy trial would *ex ante* specify outcomes that reveal whether the policy achieves those goals. For example, a trial of after-school programming designed to improve academic performance should indeed assess the short- and long-term impacts on children's academic performance. But such programs may also yield benefits through other pathways, such as providing safer after-school supervision for children while parents work outside the home. Policy-

---

159. COUNCIL ECON. ADVISERS, https://www.whitehouse.gov/cea/ (last visited Dec. 21, 2018) [https://perma.cc/M2LH-QEXC].

160. RANDOMIZED CONTROLLED TRIALS AND EVIDENCE-BASED POLICY: A MULTIDISCIPLINARY DIALOGUE, 210 SOC. SCI. & MED. 1 (Ichiro Kawachi et al. eds., 2018).

161. John P. A. Ioannidis, *Randomized Controlled Trials: Often Flawed, Mostly Useless, Clearly Indispensable: A Commentary on Deaton and Cartwright*, 210 SOC. SCI. & MED. 53 (2018).

162. Michael Sanders & Aisling Ni Chonaire, *"Powered to Detect Small Effect Sizes": You Keep Saying That. I Do Not Think It Means What You Think it Means.* (Ctr. for Market & Public Org., Working Paper No. 15/337, 2015), http://www.bris.ac.uk/media-library/sites/cmpo/documents/WP15337_Web_Version.pdf [https://perma.cc/N3Q3-T2FB].

makers may not wish to continue funding for programs on the basis of these positive externalities should they fail to improve academic performance. They may decide that those positive externalities are better served by another program or funding stream, or that they do not deserve governmental funding at all. But a failure to measure all outcomes of practical relevance, including positive and negative externalities, may cause evaluations to underestimate, or overestimate, the full benefits and costs of those governmental choices. Planning a policy trial must therefore entail groundwork to identify outcomes of relevance not only for assessing whether the program works as intended, but also for assessing whether the program has additional, perhaps unintended impacts.

   Two structural factors are also important to secure an evaluation that is as free from bias as possible: investigator independence and a priori transparency about evaluation methodology and planned outcomes. Although best-practices research methodology is objective, there is nonetheless leeway for researchers to exercise judgment when specifying hypotheses, including relevant outcomes, planning statistical tests, and managing data limitations. Ensuring that studies are done by a party without a stake in the results is an important priority for limiting bias in the exercise of those judgments. Pre-publication of evaluation methodology is also a crucial strategy to defend against statistical fishing expeditions, selective outcome reporting, and the "file drawer problem" (by which nonsignificant findings are not released).[163] Experiences in medical literature with mandatory pre-registration of clinical trials have shown that pre-publication of study methods and hypotheses leads to the reporting of smaller treatment effect estimates,[164] suggesting that the strategy provides a defense against selective outcome reporting.

---

163. Annie Franco, Neil Malhotra & Gabor Simonovits, *Publication Bias in the Social Sciences: Unlocking the File Drawer*, 345 SCI. 1502 (2014).

164. An-Wen Chan, *Bias, Spin, and Misreporting: Time for Full Access to Trial Protocols and Results*, 5 PLOS MED. 1533 (2008); Agnèn Dechartres et al., *Association Between Trial Registration and Treatment Effect Estimates: A Meta-Epidemiological Study*, 14 BMC MED. 1 (2016); Sylvain Mathieu et al., *Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials*, 30 JAMA 977 (2009); Deborah A. Zarin et al., *The ClinicalTrials.gov Results Database – Update and Key Issues*, 364 NEW ENG. J. MED. 852 (2011). *But see* Shane Killeen, *Registration Rates, Adequacy of Registration, and a Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials Published in Surgery Journals*, 259 ANNALS SURGERY 193 (2014).

There have been some improvements in researcher independence and evaluation transparency, at least in the healthcare context. Here, the experience of CMS in overseeing mandatory state-led evaluations of experimental (§ 1115) Medicaid programming is instructive. A recent report by GAO noted deep-seated methodological problems in these evaluations, including a lack of control groups, selection of inappropriate control groups, and a lack of statistical power due to small sample sizes.[165] Most evaluations prior to 2014 had been completed by the states themselves, without use of an independent outside investigator. Before the Affordable Care Act, these evaluations were also private, and evaluation reports were issued to CMS long after the window for program reapproval had passed (by which time reapproval had already been issued).[166] In 2014, however, CMS began to require states implementing experimental Medicaid programming to use an independent evaluator, and the agency began to specify in state approval letters that evaluations must use control groups and report on study limitations. The Affordable Care Act also requires prior publication of evaluation methods, suggesting that this protection may be feasible for other types of policy evaluation.

Research that evaluates legislation, regulation, and governmental programs is not alone in these limitations. These drawbacks are pervasive in much of medical research as well, and they are the subject of ongoing critique and concern.[167] Research on the effectiveness of governmental choices can therefore learn from best practices in medicine and other fields grappling with these problems. The problems I have identified are also a partial list; a host of other concerns matter in assessing policy research initiated or sponsored by the government, including questions about data infrastructure, reporting, privacy, and data linkages across agencies. But in this preliminary discussion, I have intended to raise some of the most salient ways in which evaluations can fail to identify "what works."

---

165. U.S. GOV'T ACCOUNTABILITY OFFICE, MEDICAID DEMONSTRATIONS: EVALUATIONS YIELDED LIMITED RESULTS UNDERSCORING NEED FOR CHANGES TO FEDERAL POLICIES AND PROCEDURES (2018), https://www.gao.gov/products/GAO-18-220 [https://perma.cc/3BS5-CTHJ] [hereinafter MEDICAID DEMONSTRATIONS].

166*. Id.*

167. Iain Chalmers & Paul Glasziou, *Avoidable Waste in the Production and Reporting of Research Evidence*, 374 LANCET 86 (2009).

B. Sham EBPM

Pathologies of evidence production are likely to be resolved more easily than pathologies in evidence application. Because methodological best practices exist for empirical research, and because research is generally reported in sufficient detail to understand departure from those practices, errors in research are often visible and correctable. But mistaken or disingenuous applications of empirical evidence may be harder to detect. I propose here a typology of ways in which feedback loops between empirical evidence and policy choices can break down, such that EBPM is unlikely to function in a particular policy context.

1. Technical Breakdowns

Technical breakdowns occur when regulators and legislators intend to draw on empirical evidence, but are unable to do so due to practical failures—for example, inability to find evidence, lack of time or financial resources to support research, lack of relationships with people who produce or interpret research, and lack of capacity to interpret evidence.[168] The impact of technical breakdowns may be a partial EBPM process—an attempt to use whatever evidence is available, to the extent possible—and a resulting decision that is suboptimal or ineffective, but perhaps still touted as "evidence-based." Some of these problems are intractable; as described above, the bounded rationality of policy-makers and bounded objectivity of scientific research can affect the generation and use of evidence, contributing to technical breakdowns even among good-faith EBPM proponents. But many of these limitations could be mitigated with sufficient time and resources, such as allocating funds for independent panels to synthesize research on policy questions, expanding and funding evaluation mandates, or building legislative capacity to identify and use empirical research.

Although technical breakdowns may be inadvertent, they may also reflect deliberate choices by policy-makers to spend finite resources on parsing and generating research evidence. The attention, financial support, and technical capacity needed to engage in a robust form of EBPM may be considerable, which poses opportunity costs for legislators and regulators with limited time, attention, and budgets for staff and research. Although policy-makers may support the goals of EBPM, valid concerns about the time and effort needed to sift through research evidence may undermine

---

168. *See*, *e.g.*, Oliver, *supra* note 36.

EBPM in practice. Given these constraints, "muddling through" with imperfect information may be an understandable and even optimal choice given tradeoffs with other priorities.

EBPM advocates should continue to seek ways to make engagement with research cheaper and easier, as well as ways to encourage the development of research that is immediately responsive to policy priorities.[169] Some of the solutions identified below for advancing EBPM are comparatively cheap, such as promoting evaluation transparency, making ex ante policy commitments tied to planned evaluation research, or asking for teams of rivals in independent evaluations. Other solutions are more expensive, such as enacting and funding rigorous evaluation mandates. The Foundations for Evidence-Based Policymaking Act describes a centralized, OMB-led effort to amass statistical evidence,[170] which benefits from economies of scale.

2. Evidence Distortion

A more insidious breakdown in the feedback loop between evidence and policy is the deliberate exclusion or distortion of evidence to facilitate decisions that are based on political preferences, unfettered by contrary research findings. These outcomes are likely when values are in conflict; removing choices to neutral technocratic territory is unlikely when there is dispute over what values should govern the choice.[171] When decision-makers manipulate the available evidence to support their political choices, however, they co-opt the scientific legitimacy of an "evidence-based" approach.

Here, I consider evidence distortion that is deliberate; I posit that decision-makers *know* the evidence they are suppressing, and that they understand how they are misrepresenting research findings. But these pathologies are exacerbated by the bounded rationality of decision-makers, the bounded objectivity of science, the bounded capacity of science to measure outcomes that matter (particularly value-laden concerns), and the need to preserve political participation alongside the role of research

---

169. For a call for research that is responsive to federal agency needs, see Chien, *supra* note 155.

170*. See* Foundations for Evidence-Based Policymaking Act of 2017, Pub. L. No. 115-435, 132 Stat. 5529 (2019).

171*. See*, *e.g.*, Claire Huntington, *The Empirical Turn in Family Law*, 118 COLUM. L. REV. 227 (2018) (arguing for clarity on which values should govern a choice, before the turn to empirical research).

evidence. Regulators with bounded rationality may believe that they are seeing the evidence clearly, whereas their comprehension of research is instead shaped by cultural concerns and predictable cognitive biases. Political participants in policy development may also exaggerate concerns about bounded scientific objectivity, bounded capacity to measure meaningful outcomes, and the bounded authority of science to dismiss or halt unwelcome but valid studies (or to elevate research findings that align with their political norms). Purposeful, disingenuous evidence distortion is thus its own form of breakdown, but the inherent limitations of EBPM can exacerbate this problem.

A few examples may be instructive. Upward Bound is a federally funded program intended to assist poor students prepare for college. An independent experimental evaluation of Upward Bound began in 1991, funded by the U.S. Department of Education. The results of this study were disappointing, suggesting that although students were more likely to acquire high school math credit and earn a vocational license, there was no detectable effect on graduation rates, grades, college enrollment, or college degree completion.[172] A second randomized study of the program was intended to begin in the late 2000s, around the time that the first round of evaluation reports was released. Congress, however, barred the use of federal funds for this evaluation in its 2008 budget appropriations,[173] and the Higher Education Act of 2008 stipulated that future evaluations should focus on "identify[ing] . . . project practices that are effective"; moreover, the Secretary could not require Upward Bound grantees to agree to an evaluation that "requires the eligible entity to recruit additional students . . . ; or results in the denial of services for an eligible student."[174] These actions effectively barred the use of randomized assignment in future Upward Bound evaluations.[175] Rather than expose the project to further study, Congress halted the creation of evidence that might interfere with the reauthorization of Upward Bound funds.

Evaluations of a federal school voucher program in Washington, D.C. met a similar fate. In the spring of 2017, a randomized study identified harmful results of the federal voucher program for students who received

---

172. *National Evaluations of Upward Bound*, MATHEMATICAL POL'Y RESEARCH, https://www.mathematica-mpr.com/our-publications-and-findings/projects/upward-bound [https://perma.cc/Q8NQ-MV6E].

173. Viadero, *supra* note 111.

174. Higher Education Opportunity Act, Pub. L. No. 110-315, 122 Stat. 3078 (2008).

175. *See* Viadero, *supra* note 111.

vouchers, compared to applicants who did not.[176] A week later, Congress proposed legislation that barred the use of randomized controlled trials to evaluate the program in future years.[177] Again, the legislation was justified in terms of "not deny[ing] scholarships to certain students placed into a control group." But the timing of this change was no coincidence, given that the initial evaluation results were contrary to the political preferences of legislators promoting the program. The program has been reauthorized through 2019 with an appropriation of $17.5 million.[178]

Congress is not the only set of policy makers to suppress unfavorable evidence. In recent months, many have noted actions taken by the EPA under the direction of administrator Scott Pruitt. In rules issued in the fall of 2017, the EPA barred those who receive agency grant money from serving on panels that advise the agency on regulatory decisions—a move that precluded many academic researchers from providing scientific expertise on these panels.[179] In the aftermath of the new rule, the number of academic researchers on boards fell by fifty percent, while the number of board members from regulated industries tripled.[180] Before his departure from the agency, Pruitt issued a set of proposed regulations for notice and

---

176. MARK DYNARSKI ET AL., EVALUATION OF THE DC OPPORTUNITY SCHOLARSHIP PROGRAM: IMPACTS AFTER ONE YEAR (U.S. Department of Education, 2017), available at https://ies.ed.gov/ncee/pubs/20174022/pdf/20174022.pdf [https://perma.cc/SS4Y-LJLF].

177. Sarah D. Sparks, *Congress Budget Deal Bans New Gold-Standard Studies of Federal Vouchers*, EDUC. WEEK BLOGS (May 2, 2017), https://blogs.edweek.org/edweek/inside-school-research/2017/05/congress_budget_deal_bans_new.html [https://perma.cc/9FVC-H6VY]; *see also* Gamoran, *supra* note 35.

178. *See* Congressional Research Service, *District of Columbia Opportunity Scholarship Program (DC OSP): Overview, Implementation, and Issues* 1 (2019), https://fas.org/sgp/crs/misc/R45581.pdf [https://perma.cc/HZ8C-7M7J].

179. EPA, *Administrator Pruitt Issues Directive to Ensure Independence, Geographic Diversity & Integrity in EPA Science Committees* (Oct. 31, 2017), https://www.epa.gov/newsreleases/administrator-pruitt-issues-directive-ensure-independence-geographic-diversity [https://perma.cc/V9MF-5V96].

180. *Swamp Science: Scott Pruitt Embarks on a Campaign to Stifle Science at the EPA*, THE ECONOMIST (Apr. 26, 2018), https://www.economist.com/united-states/2018/04/26/scott-pruitt-embarks-on-a-campaign-to-stifle-science-at-the-epa [https://perma.cc/U5KR-4Z7J].

comment, entitled "Strengthening Transparency in Regulatory Science."[181] If enacted, the regulations will limit the agency's ability to consider scientific studies that do not make the underlying data publicly available—which is impossible given the human subjects protections in place for many academic research studies, particularly those using datasets like medical records.[182] Proposed legislation in Congress would have taken similar steps if successful.[183]

These are visible examples, set out in agency orders or legislation. More subtle exclusion of evidence is achieved by simply not considering it, weighing unfavorable evidence lightly, or subjecting favorable vs. unfavorable evidence to differing levels of methodological scrutiny.

Empirical evidence cannot resolve every conflict. Some decisions cannot be made without first resolving thoroughgoing conflicts of values, because there is little agreement about what question evidence should answer. For instance, empirical evidence is surely relevant to the question of whether medical providers who perform abortions not necessary to save the life of the mother should be eligible to participate in Medicaid. There is evidence pertaining to the other medical services those providers perform, the quality of those services, and the likely effects on access to healthcare of excluding those providers from Medicaid reimbursement. But if history is any guide, policy-makers' choices about whether to cover these providers will not draw on this evidence.

Why won't evidence help? The dispute is ideological—we are, as policy scholar Roger Pielke, Jr. would say, in the realm of "abortion politics," not the realm of "tornado politics." In abortion politics, information is used to rationalize, to help justify decision commitments already made, to provide

---

181. Strengthening Transparency in Regulatory Science, 83 Fed. Reg. 18768 (proposed Apr. 20, 2018) (to be codified at 40 C.F.C. pt. 30).

182. Kathleen Rest & Georges C. Benjamin, *Trump's EPA Puts Our Health at Risk*, Sᴄɪ. Aᴍ. (Jul. 13, 2018), https://blogs.scientificamerican.com/observations/trumps-epa-puts-our-health-at-risk [https://perma.cc/D34S-884W]; Jeremy Berg et al., Letters, *Joint Statement on EPA Proposed Rule and Public Availability of Data*, 360 Sᴄɪ. (May 4, 2018), http://science.sciencemag.org/content/360/6388/eaau0116 [https://perma.cc/7UU3-BE3F]; Alaska Community Action on Toxics et al., Comment Letter on Inadequate Opportunity for Public Engagement— Strengthening Transparency in Regulatory Science (June 20, 2018), https://www.regulations.gov/document?D=EPA-HQ-OA-2018-0259-3644 [https://perma.cc/7VSP-DE8M].

183. HONEST Act, H.R. 1430, 115th Cong. (2017); *see also Swamp Science*, *supra* note 180.

narrative, and to shore up power. In tornado politics, information is used to assess decision alternatives, to evaluate choices, to deduce possibilities, and to make decisions.[184] Abortion politics extends to many realms—consider, for example, debates over gun control, the death penalty, climate change, immigration, and nondiscrimination law. Tornado politics, by contrast, is comparatively anodyne, with current examples perhaps extending to funding for opioid abuse treatment, infrastructure improvements, and defense spending—all of which have given rise to bills passed with bipartisan support in recent years.

Evidence may indeed have its uses in abortion politics, but enough repetition should convince us that this evidence is doing little to inform the choices being made. And indeed, empirical evidence may make the difficult values-based conversations about abortion more difficult.[185] As Pielke notes, "Conflation, often willful, of Abortion and Tornado Politics encourages the mapping of established interests from across the political spectrum onto science and then uses science as a proxy for political battle over those interests."[186] This may be inevitable given the bounded rationality limitations noted above, and, in fact, there is reason to believe that importing scientific evidence into these conversations will make them more polarized.[187] We should not, however, delude ourselves into believing that this is "evidence-based" policy. And we should not overlook the harm that the misuse of this label may pose to policies that do derive from empirical evidence.

3. Terminal Experimentation

Another interruption in the feedback loop between evidence and policy is the cultivation of uncertainty in order to justify using experimental policies as a stand-in for more decisive action—either as a substitute for doing nothing, or as a substitute for doing something more lasting and impactful. In either of these situations, there may not be political momentum to achieve a preferred policy goal—some may wish to make a

---

184. ROGER A. PIELKE, THE HONEST BROKER: MAKING SENSE OF SCIENCE IN POLICY AND POLITICS 40–43 (2007).

185. *Id.* at 47.

186. *Id.* (noting that "scientific information . . . would represent a distraction from the task of reconciling different value commitments through bargaining, negotiation, and compromise").

187. DAN M. KAHAN ET AL., AFFECT, VALUES, AND NANOTECHNOLOGY RISK PERCEPTIONS: AN EXPERIMENTAL INVESTIGATION (2007).

long-term financial investment in a form of social programming, for example, but there is insufficient political will to do so. Or, conversely, some may wish to abolish a category of social programming entirely, but they may be unable to persuade others to do so. From either direction, classifying activities as "experimental" and amplifying uncertainty about long-term effects is one means of achieving political buy-in. It is easier to agree to a time-limited experiment with finite resources and commitment, compared to supporting a full-fledged course of policy action that commits a larger set of resources and subjects a larger proportion of the population to the intervention. Asking others to buy into an experiment also benefits from the halo of scientific legitimacy. Even if a program is itself of uncertain effectiveness, attaching an evaluation connotes good faith on the part of the experimenter, and it implicitly promises to revisit the policy choice when the experiment concludes.

All of this may be acceptable and even salutary in the short term, especially when such efforts are accompanied by genuine and well-resourced evaluations. The results of those evaluations can indeed inform policy choices, including longer-term investments and codifications of policy choices in law. In the short term, this is EBPM working as intended.

But in the long term, there is a danger that the evolution from pilot study to more stable policy (including scaling up or scaling down, as warranted) will not occur. When there is strong momentum against a longer-term policy, continued experimentation becomes "symbolic"—it is pathological when "experiments are conducted to replace real action with symbolic action."[188] Typically, this is understood to be a move in a conservative direction.[189] Namely, emphasizing uncertainty about program impacts would justify small-scale experimentation as permanent stasis, rather than proceeding to larger-scale programming and a stable funding stream. It need not be thus, however; we could imagine governmental programming that does not work and is unambiguously unsupported by evidence. Terminal experimentation would exaggerate scientific uncertainty to justify the continuation of such programming as a pilot or time-limited choice, rather than using evidence to eliminate it.

As with evidence distortion, the pathology of terminal experimentation has plenty of help from some of the inherent boundaries on EBPM. Normative, cultural, and political commitments may make policy-makers skeptical of programmatic choices, which will contribute to skepticism regarding study findings. Those who seek to hold programs in long-term

---

188. GREENBERG ET AL., *supra* note 83, at 47.

189*. Id.* (calling symbolic experimentation an "inherently conservative force").

limbo might also draw on the bounded objectivity and authority of science to impugn the findings of past research and to call for new studies with minor tweaks. Those who are unsatisfied with policy choices on grounds that are harder to measure, such as concerns about dignity or distributional effects, are unlikely to be convinced by research studies that do not account for these concerns. Like evidence distortion, terminal experimentation fundamentally suggests a deliberate choice—here, the choice to replace action with indefinite, symbolic experimentation. But the inherent limitations on EBPM furnish ammunition that amplifies the problem.

The line between EBPM as intended and terminal experimentation is difficult to fix, but it invokes questions of intent. In terminal experimentation, policy-makers deliberately seek to exaggerate scientific uncertainty in order to avoid long-term programmatic choices. If the intent of the decision-makers is to avoid the entrenchment of a program (or, conversely, to maintain a program despite lackluster evaluations), terminal experimentation is likely. For example, David Super's research on safety-net programming provides a compelling argument that an experimental approach in welfare policy has served not only to undermine the long-term stability of welfare programs, but also to allow capture of state welfare programming by adverse political interests.[190] On this view, experimentation is not just a means of avoiding more permanent policies, but it can also provide a platform for undermining policy goals. This takes advantage of the same dynamic mentioned above—it is easier to secure agreement to change that is "experimental" in nature, rather than a permanent policy change, and the entrenchment of ongoing experiments as a mode of policy operation can allow long-term acquiescence to those changes.

4. Ratcheting

A final interruption in the link between evidence and policy is the deliberate misuse of evidence for ratcheting. This reflects the selective use of empirical evidence to mask a unilateral, values-driven project: evidence is consulted, but only the findings that support a preexisting political goal are used, and those findings are applied far more broadly than is warranted. The conclusions of this sham EBPM process are thus predetermined by politics, and they reach further than the evidence would support. The problem of ratcheting escalates when there is a lack of consensus on which

---

190. David Super, *Laboratories of Destitution: Democratic Experimentalism and the Failure of Antipoverty Law*, 157 U. Pa. L. Rev. 541 (2008).

outcomes matter for determining that a program "works." Many programs do not work, or they work for some outcomes and not others, and the outcomes we care about will depend on our policy preferences. When the effect of a program is mixed—there are benefits for some outcomes but not others—a motivated decision-maker relies only on selected outcomes to justify a wholesale decision to discard (or, theoretically, scale up) funding for the category of programming.

The use of EBPM for ratcheting is typically associated with dismantling social programming, and efforts by Republican administrations to promote policy evaluation have stoked fears that the results will be used purely for funding cuts.[191] Deborah Stack has described this dynamic among agency administrators reacting to the Bush-era PART program: "[B]ecause rigorous evaluations of whole programs usually produced null findings, agencies viewed evaluation as a threat rather than as a tool for program improvement. They feared that Republicans might use null findings as a reason to cut program funding."[192] Indeed, analyses have found that under the Bush administration, PART scores were more likely to result in proposed budget cuts for programs created under a Democratic administration. "Put more simply, liberal programs were exposed to the risks of performance budgeting, while conservative programs were not."[193] The converse may, however, also be true if administrators were to selectively use empirical results (say, findings for secondary or irrelevant outcomes) to expand and entrench programs that made little difference on outcomes of importance.

Where individual policy efforts do not perform well, there is little argument to continue funding them without change. Decommissioning programs that are ineffective is an important part of the efficiency rationale for EBPM. But the concern with ratcheting is that instead of just de-funding an ineffective *approach* to a policy goal (e.g., Upward Bound for promoting high school graduation, D.A.R.E. for reducing substance use), policy-makers reacting to negative or null results will de-fund the entire *policy goal* without reallocating resources to other means of achieving the goal. Say, for example, that D.A.R.E. has proven ineffective for reducing drug use among adolescents.[194] An administrator or legislator engaging in ratcheting would

---

191. Stack, *supra* note 74.

192*. Id.*

193. Moynihan, *supra* note 98, at 14.

194. Indeed, it is ineffective, as generations of evaluations have shown. Steven L. West & Keri K. O'Neal, *Project D.A.R.E. Outcome Effectiveness Revisited*, 94 AM. J. PUB. HEALTH 1027 (2004).

use the finding to eliminate funding for the *entire category* of youth substance abuse prevention—a shift in *priority* rather than just a shift in programming. Instead, responses that avoid ratcheting would include testing efforts to reform the program or reallocating those funds to different approaches for achieving substance use prevention.

As with the other forms of sham EBPM, the tactical use of evidence for ratcheting is exacerbated by some of EBPM's inherent limitations. Boundedly rational policy-makers are likely to find studies more credible if the results align with their political priorities, and they are more likely to dismiss studies that do not as flawed. They can call upon concerns about scientific objectivity, unquantifiable concerns, and the concerns of other political actors to dismiss studies or results that do not support their preexisting normative project. Intentional ratcheting is thus a deliberate choice, but it again gains momentum from other limitations that make EBPM difficult.

The Trump administration has provided several examples of how program evaluations may be used for ratcheting. President Trump's proposed budget in 2017 included $3 billion in cuts to community development grants, as well as $2 billion in cuts to after-school programming. Budget Director Mick Mulvaney justified the cuts in the language of evidence: the programs were "not showing any results" or were lacking in rigorous evaluations.[195] In fact, the community development grants included funding for Meals on Wheels, which does have health benefits; the program also included many other types of grants to over 1,200 local governments.[196] Rather than reallocate the funding to more effective program options, however, the budget abolished it entirely. Two longtime advocates of EBPM have identified why these cuts were an abuse of the language of EBPM—namely that they free-ride on scientific legitimacy to change policy priorities:

> [E]vidence can only go so far. . . . [N]o policy evaluation can tell us the right amount to invest in the endeavors of helping kids learn or helping parents manage their lives . . . . And yet cut these efforts is precisely what the Trump budget does. It reduces education spending by $9 billion . . . . That is a reflection of values. Program evaluation is of no relevance.[197]

---

195. Gordon & Haskins, *supra* note 126.

196*. Id*.

197*. Id.*

In other words, program evaluations may have justified reallocating funding to another type of after-school programming, particularly if there was consensus that the only purpose of the funding was to promote academic achievement and not to support working families. But program evaluations would not have justified ratcheting: de-funding an entire category of social programming on the basis of selective results.

Another example arises from the field of teen pregnancy prevention. Under the Teen Pregnancy Prevention Program (TPPP), created by the Obama administration, HHS awarded federal grants to 84 organizations to implement pregnancy prevention programming, with an emphasis on services for teens of color. Programs were required to include impact evaluations. In July of 2017, HHS gave notice to all funded programs that their funds were rescinded; instead of being funded through 2020, the programs would be terminated after another year, and the president's budget proposed eliminating the program entirely. HHS justified these funding cuts on the basis of evidence: in the words of an HHS spokesperson, "Of the 37 projects funded and evaluated for a 2016 report, 73 percent had no impact or had a negative impact on behavior."[198] Some of these grantees have filed suit, and at least five have now won preliminary motions in federal court.[199] In the meantime, the administration has released funding opportunity announcements for programming that *does not* use one of 48 programs previously determined by HHS to be effective, but instead selects one of two approaches that promote sexual abstinence.[200]

These are cuts by ratcheting: eliminating an entire category of programming, despite the fact that some eliminated activities are beneficial or of unknown effectiveness. The redirection of funds to other methods of achieving the same program goal is less in keeping with ratcheting. But here, peer-reviewed literature has for decades found that abstinence-only

---

198. Alison Kodjak, *Trump Administration Sued over Ending Funding of Teen Pregnancy Programs*, NPR (Feb. 15, 2018), https://www.npr.org/ sections/health-shots/2018/02/15/585879601/trump-administration-sued-for-ending-funding-of-teen-pregnancy-programs [https://perma.cc/8DZX-QEYW].

199. Jessie Hellmann, *Judge Orders Trump Officials to Process Grants for Teen Pregnancy Programs*, THE HILL (June 4, 2018), http://thehill.com/policy/healthcare/390572-court-sides-with-grantees-in-class-action-lawsuit-over-teen-pregnancy [https://perma.cc/3U5Z-4SK6].

200. Jessica Boyer, *The Teen Pregnancy Prevention Program Was on the Right Track, Now It's Being Dismantled*, HEALTH AFF. BLOG (May 24, 2018), https://www.healthaffairs.org/do/10.1377/hblog20180523.207226/full [https://perma.cc/4QKM-H3SS].

programming is ineffective for achieving the stated goal of pregnancy prevention,[201] the standards used here to appraise evidence for the two types of programming are thoroughly different, and no observer familiar with this evidence base would agree that this is EBPM working as intended.

Yet another example comes from recent HHS cuts to the budget for "navigators": nonprofit organizations that help sign people up for health insurance on the exchanges. The statutory purpose of navigators is to fulfill several functions: namely, they must "perform public education and outreach activities; distribute fair and impartial enrollment information on health plans and the availability of federal subsidies; facilitate enrollment in qualified health plans; provide referrals to appropriate agencies for grievances or complaints; and provide all information in a manner that is linguistically and culturally appropriate for the consumer."[202] Evaluations thus far have suggested that navigators do, in fact, prompt insurance signups.[203] But under the Trump administration, HHS has cut navigator funds by more than eighty percent total, framing the decision as justified by evidence that navigators sign up fewer people than insurance agents and brokers (who differ from navigators because they receive commissions on the basis of the products they sell; navigators are barred from doing so).[204] HHS officials announced seventeen navigator programs that had signed up one-hundred people or fewer, but rather than reallocating funds to other navigator programs, the entire navigator budget was cut.

Without mechanisms in place to preserve funding allocations—such that ineffective approaches within funding categories are discontinued, but that funds are reallocated to program reform or effective approaches within

---

201. Underhill et al., *supra* note 127.

202. Sarah Goodell et al., *Health Policy Brief: Navigators and Assisters in the Third Open Enrollment Period*, HEALTH AFF. (Sept. 28, 2015), https://www.healthaffairs.org/do/10.1377/hpb20150928.132418/full [https://perma.cc/VCE6-XDVS].

203. Benjamin D. Sommers et al., *The Impact of State Policies on ACA Applications and Enrollment Among Low-Income Adults in Arkansas, Kentucky, and Texas*, 34 HEALTH AFF. 1010 (2015); Goodell et al., *supra* note 202; Michele Thornton, The Role of External Support in Health Insurance Decision-Making Under the Affordable Care Act (Jan. 29, 2018) (unpublished Ph.D. thesis, University of Illinois at Chicago), https://indigo.uic.edu/handle/10027/22603 [https://perma.cc/3E4M-7F3G].

204. Robert Pear, *Trump Officials Slash Grants that Help Consumers Get Obamacare*, N.Y. TIMES (July 10, 2018), https://www.nytimes.com/2018/07/10/us/politics/trump-affordable-care-act.html [https://perma.cc/H5NA-SL9S].

the same category—the fear of ratcheting will attend any serious calls for program evaluation, especially under conservative administrations. This can lead to gaming, avoidance of rigorous evaluation methods, and reluctance to provide the evidence that would be useful for long-term policy choices.

III. THE COSTS (AND BENEFITS) OF SHAM EBPM

When the uses and generation of evidence veer so far from EBPM as to be "sham," there are adverse consequences measured not only against ideal policies (policy choices miss opportunities to be more effective), but also against the status quo (policy choices inflict new harm). Herein I will distinguish between broken experimentation and sham application of evidence, and I consider numerous ways in which sham applications of evidence can be harmful. These include wasting resources on ineffective policies, free-riding on the credibility of "evidence-based" terminology, and undermining scientific legitimacy. Some of these are also hazards of broken experimentation, which poses additional harms to human subjects. Flimsy experimentation, however, may also pose some positive externalities— even when evidence does not inform primary policy decisions as intended, rigorous evaluations may nonetheless inform choices in other policy and practice settings (collateral EBPM), improve policy implementation by agencies who are aware that they are being evaluated (the Hawthorne effect), and produce evidence that clarifies the stakes of values-based reasoning even if it does not drive the policy choice. The goal of this Part is to identify the costs and inadvertent benefits of sham EBPM processes, with an eye toward proposals that limit or minimize these harms.

A. Welfare and Efficiency

It is easiest to see the threats that broken experimentation and sham EBPM pose to welfare and efficiency. Halfhearted or subversive flaws in the generation and use of evidence can waste governmental resources, or, in the case of ratcheting, result in the wholesale cutting of categories of programming that serve valuable social ends. Moreover, sham EBPM that proceeds to neglect or distort evidence can result in outcomes that are either inefficient or harmful in relation to the stated policy goals.

To consider broken experimentation first, it is clear at the outset that rigorous evaluations can be expensive. A methodologically strong evaluation will need to cover independent evaluators, the generation of the study protocol, data collection and management, data analysis, and public

reporting, among other costs. Appropriating funds for evaluation may cut into program budgets, and lack of resources is a principal reason for poor-quality evaluations in programs such as Medicaid Section 1115 waivers.[205] But *failing* to spend sufficient funds for high-quality evaluation can paradoxically turn well-intended evaluation dollars into waste. Spending too little can lead to a flimsy design, such as a pre-post evaluation that lacks a control group, that fails to measure important outcomes, or that lacks independence. When evaluations are fatally flawed in their methodology—when they are broken—evaluation funds are wasted resources.

Policy evaluations are not the only subset of research that gives rise to wasted experimentation funds; indeed, some of the top medical journals have recently dedicated space to soul-searching about the utility of clinical research.[206] At the outset of a powerful series on waste in medical research published by the Lancet, evidence-based medicine advocate Iain Chalmers identifies waste arising from breakdowns at every set of the research process: (1) failures to ask questions that are high-priority and novel; (2) failures to use appropriate research designs, methods, and analysis; (3) inadequate or costly compliance with research oversight burdens; (4) failures to report methodology and outcomes of interest; and (5) provision of research reports that are biased or unusable in practice.[207] Although it is a strength to include experimentation and appropriate evaluation funds in government programming, poorly planned and poorly executed evaluations are subject to the same categories of waste. Evaluations subject to the sources of bias discussed above are unable to contribute results of relevance for future decision-making, and the results of poorly executed evaluations are likely to be misleading, leading to costly errors in future decisions.

Sham EBPM is likely to be even more costly. The different categories of sham EBPM (technical breakdowns, evidence distortion, terminal experimentation, and ratcheting) have different origins and somewhat different effects. But each of them leads to a similar problem: inefficient allocation of governmental resources, with adverse consequences for both expenditures and public impact of government choices. Simply, decision-making procedures are unlikely to reach good results when they omit evidence, neglect evidence selectively, play up uncertainty to do nothing, or use evidence disingenuously to ratchet according to policy preferences. In

---

205. U.S. Gov't Accountability Office, Medicaid Demonstrations, *supra* note 165.

206. Chalmers & Glasziou, *supra* note 167; John P. A. Ioannidis et al., *Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis*, 383 Lancet 166 (2014).

207. Chalmers & Glasziou, *supra* note 167.

the three prior categories, inefficiency is likely the result of evidence that might points toward more effective or efficient means of accomplishing governmental goals being neglected. In the case of ratcheting, the uses of evidence to cut programming leads to reduced governmental expenditures, but the loss of effective programming may lead to outsized public harms; the uses of evidence to expand ineffective programming leads to increased expense with little benefit.

Of course, "inefficient" decisions may sometimes be correct when they reflect values that supersede cost-benefit concerns.[208] Here, however, it matters that evidence-based decisions take as their starting premise the purpose of the policy: given the goal of the policy or program, what is the most effective choice? Competing values properly factor into the *goal*, and not simply the cost-benefit calculus in comparing different policy options. Choosing a more expensive program for, say, dignitary reasons (as above) is not a rationale for an inefficient decision—it is instead a means of locating the most effective choice to fulfill policy goals *including* the promotion of dignity. These values-based choices, consequently, are not departures from EBPM. Decisions that neglect or distort empirical evidence of benefit and harm, however, are troubling because they are more likely to lead to fruitless spending, missed opportunities to improve people's lives, and sometimes outright harm.

B. Shallow Signaling and Scientific Legitimacy

In addition to undermining policy ends, broken experimentation and sham EBPM also benefit unfairly from free-riding on scientific legitimacy. The misuse of evidence in policy, however, can contribute to the erosion of that legitimacy over time. Evidence distortion and the erosion of scientific legitimacy are distinct harms—the former amplifies the inefficiency and social hazards of policies that neglect evidence, while the latter inflicts independent injury on public confidence in empirical evidence and the utility of science. Ultimately, the instrumental effects of lost public confidence in empirical evidence are likely to be *more* waste and missed opportunities for promoting the public good.

---

208. *Supra* Section I.B (describing how values such as dignity and equity can elude cost-benefit analysis).

1. Free-Riding

Beliefs about the value and purpose of science, including social science, are largely positive. In a 2016 national survey, public trust in science and scientists was strong and has been fairly stable since the 1970s.[209] When an agency administrator or legislator seeks to persuade others of the virtues of a particular policy choice, they can free-ride on scientific legitimacy by using the language of scientific evidence. EBPM includes two types of communications: describing the empirical evidence for a decision, and offering to subject the program or policy to an evaluation. Both communications are expressive, and both communicate more than simply the fact of evidence, or the fact of wanting to do an evaluation. Given popular beliefs about the legitimacy and intentions of empirical research, these statements also reveal information about the intentions, good faith, and political posture of the legislator or administrator herself, and these are likely to generate political support for her position. When the legislator or administrator is in fact practicing sham EBPM, however, she is free-riding on scientific legitimacy—the legitimacy halo is undeserved.

Communicating the empirical evidence underlying a policy choice is in part an expression of the values of the decision-maker. By engaging with evidence, the decision-maker demonstrates to others that she is not purely concerned with political ends—she values the impact of the decision on the public, and (impliedly) she would adjust her choice if the evidence were different. She communicates that she values the goal of the policy as against other ends, and she also signals in part that the policy is contingent on its continued performance. She communicates that the policy is feasible, or at least sufficiently reasonable to have been tested before, and a good-faith belief that the policy will work as intended. She also communicates a set of personal characteristics that includes open-mindedness to others' knowledge (albeit here, limited to experts), which is politically desirable for policy-makers expected to serve a constituency. Finally, she implies her own personal ability to understand the methods and results of the studies she cites.

---

209. Trust in science on particular issues, however, varies depending on people's prior beliefs on issues such as vaccines, climate change, and genetically modified foods; on each of these issues, scientists were rated as somewhat less trustworthy. Cary Funk, *Mixed Messages About Public Trust in Science*, PEW RES. CTR. (Dec. 8, 2017), http://www.pewinternet.org/2017/12/08/mixed-messages-about-public-trust-in-science [https://perma.cc/4DDK-EUKX].

Scientific legitimacy also extends to the intention to evaluate one's choices, and policy-makers who seek to subject their own programs to evaluation are similarly communicating with both instrumental and expressive effect. Taking an agnostic view of the policy ends, the act of evaluation communicates information about the decision-maker's priorities, her views of the participants subject to the experiment, her commitment and good-faith rationale for selecting the policy under study, and her tentative commitment to the policy over time. In other words, proposing experimentation sends signals to policy advocates and opponents, the individuals subject to the provisional rule or program, and current and future administrations responsible for implementing and continuing to enforce the new rules. Evaluating a program also signals that it is temporary, or contingent, and therefore less threatening to opponents than a permanent policy choice. Opponents will anticipate an opportunity, that is, to parse the evidence and to argue against the program on empirical grounds later, and this expected opportunity can lower the stakes of the current debate. Even when the evaluation is purely an idea, the policy position may gain a legitimacy bump when its sponsors propose an attendant evaluation.

These signals are politically useful. Popular beliefs about the value and virtue of science can help policy-makers achieve consensus on their ideas—either because those ideas are backed by science or because they will be subjected to an evaluation later. In the case of broken experimentation and sham EBPM, however, these signals are shallow if not entirely hollow, and they cultivate political capital and credibility where it should be absent. This can build political momentum for (or decrease opposition against) policies that are wasteful, ineffective, or harmful, amplifying these harms.

2. Erosion of Scientific Legitimacy

Broken experimentation and sham EBPM are parasitic on scientific integrity, and in true parasitic form, they also act to undermine it. Public observers are not privy to the full scope of empirical evidence available for policy choices, and most observers do not have the time and resources needed to obtain and appraise the evidence themselves. People who experience the ineffective or harmful effects of policies touted as "evidence-based" will have reason to suspect that the science is biased in a way that neglects their interests or unreliable at best. Over time, people unhappy with "evidence-based" programming may come to doubt either the credibility of science itself, or the credibility of the legislative and agency

actors who purport to use evidence in their decisions.[210] Even for those who agree that empirical research generally is valuable, uncertainty about the credibility of the decision-makers who cite evidence may lead to skepticism of all "science-based" claims as disingenuous efforts to free-ride on scientific authority. This can depress the value of appeals to science in general, making it more difficult to practice genuine evidence-based practice. Where sham EBPM abounds, people are inundated with empty or contradicting claims about science, and they may come to disengage from this form of persuasion. Note that again, this is true even where people retain a basic trust in science—if they distrust the ability of decision-makers to deploy science accurately, then appeals based on scientific legitimacy will be dulled, making it more difficult for legitimate EBPM (and legitimate uses of science) to gain needed support.

C. Chilling Scientific Production of Knowledge

Sham EBPM—particularly evidence distortion and ratcheting—can also have a chilling effect on the production of scientific knowledge. Researchers grow invested in the well-being of the populations they study, or enter research in the hopes of improving outcomes for a group or community. Ratcheting poses long-term threats to funding priorities that work to benefit communities (most often policies for welfare and safety net programming), and researchers who fear ratcheting may be unwilling to expose negative or null program effects for fear of undermining all resources.

Researchers may seek to guard against this; for example, a recent book on state-by-state inequalities in Medicaid programming contains a disclaimer urging, in italics, that nothing in the book should be construed to support cuts to Medicaid programming.[211] Researchers of the Earned Income Tax Credit perceive numerous ways in which the program is too limited to support populations who are very poor,[212] but we might hesitate

---

210. Of course, if people only believe the evidence that supports their own position, then they will be suspicious of invocations of evidence any time a policy decision runs counter to their values.

211. JAMILA MICHENER, FRAGMENTED DEMOCRACY: MEDICAID, FEDERALISM, AND UNEQUAL POLITICS (2018).

212. *See, e.g.*, Anne Alstott, *Why the EITC Doesn't Make Work Pay*, 73 L. & CONTEMP. PROBS. 285, 288 (2010) (describing limitations of the EITC program, but also noting that "progressive praise for the EITC may reflect the political

to call for its repeal if it will not be replaced with a more effective solution. But besides communicating their reservations about how their research should (and should not) be used, researchers have very little control of how their results will be distorted or used in decision-making.[213] It may be no accident, therefore, that evaluations of meaningful governmental programs shy away from methodologically rigorous studies that provide robust evidence of effects. Separating budget allocations on *priorities* from allocations on *specific program approaches* could help to avoid these perverse incentives.

D. Governmental Citizens as Human Subjects

Broken experimentation and sham EBPM pose another set of hazards, namely concerns about the ethics of enrolling subjects in futile evaluations. When government-led studies have little payoff—either because they are poorly designed or because the results are highly likely to be neglected or misused in practice—continuing these studies raises ethical questions about enrolling human subjects. Program and policy evaluations themselves (as distinct from policy choices made on the basis of that evidence) may not make subjects worse off than they were before enrolling. But evaluations do burden subjects, and if these burdens cannot be justified by social benefit, these studies may be ethically suspect. This is of particular concern when the experimenter is in fact the government, which exerts multiple forms of coercive power over citizens, exists primarily to promote the long-term well-being and flourishing of its citizens, and enjoys exemptions from some research regulations.[214]

---

assumption that a modest wage subsidy is the best that U.S. politics can produce for low-wage workers").

213. "On matters of deep controversy, research findings are more likely to be mobilised as arrows in the battle of ideas, and sometimes in ways that the original authors may find distasteful." Head, *supra* note 23, at 84.

214. Under a change made during the Reagan administration, research subject to an agency head that is designed to evaluate public benefits programs is exempt from IRB review. To What Does this Policy Apply, 45 C.F.R. § 46.101 (2018); Sara Rosenbaum, *Weakening Medicaid from Within*, THE AM. PROSPECT (Oct. 19, 2017), http://prospect.org/article/weakening-medicaid-within [https://perma.cc/7FAQ-L2HH].

1. Research Lacking Practical Benefits

Research that lacks a practical benefit is categorically unethical under many sources of guidance for research ethics, including the Nuremburg Code,[215] the Council for International Organizations of Medical Sciences guidelines,[216] the Declaration of Helsinki,[217] and the Belmont Report (the primary source of guidance in the US).[218] The rationale for this is threefold.

First, research of little benefit is more likely to provide an inequitable balance of benefits and burdens. The Belmont Report (and other guidance documents, including the set of federal regulations known as the Common Rule) requires a balancing of the benefits and burdens of research. Participation imposes burdens, even if they are extremely minimal (e.g., a very small risk of a loss of confidentiality, temporary discomfort, or time). Where researchers know in advance that research poses no benefit due to sham EBPM—because it will be misused or neglected, or because it is so deficiently designed that it cannot provide reliable data—then the burdens are likely to outweigh the benefits, tipping the balance closer to unethical research.

Second, individuals who enroll in research (where they do so actively, as compared to simply being followed through administrative data) expect that the study will contribute to knowledge. They may participate for other reasons (e.g., financial incentives, a misconception that research is in fact treatment, or in-kind benefits of participation), but the expectation of contributing to knowledge is a constant, and it becomes part of the "covenant" that researchers and subjects make in the process of informed

---

215. *The Nuremberg Code (1947)*, 313 BRIT. MED. J. 1448 (1996).

216. Council for International Organizations of Medical Sciences (CIOMS), International Ethical Guidelines for Health-Related Research Involving Humans (2016).

217. World Medical Association, Declaration of Helsinki (2013), available at https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects [https://perma.cc/YW7D-2VEP].

218. NAT'L COMM'N FOR THE PROTECTION OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RES., THE BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH (1978), https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html [https://perma.cc/BBP6-L66C].

consent.[219] Where researchers fail to publish or disseminate their results—either because they choose not to do so, or because the results are somehow barred from release—they fail to fulfill the promise that subjects' study participation will contribute to knowledge, thereby undoing the bargain they made with participants.[220] It is of course possible that researchers may publish their findings in good faith, but those findings may be neglected or misused in subsequent policy choices. This happens all the time. But where it is known *in advance* that the evidence is likely to be distorted, misused, or silenced, or used for ratcheting, it complicates the ethical balancing and should factor into the disclosures that researchers make to subjects in these studies.

Third, the barrier to research that lacks social benefit reflects a concern about policing the boundaries of ethically (and therefore legally) acceptable research. Even if research subjects *knew* that the results of studies were unlikely to make an impact, or that the study was designed poorly, they may yet consent to take part in the project. But experimentation—particularly medical and governmental research[221]—has an ugly history of abuses in the U.S., in service of not only knowledge, but also the prestige and position of researchers. Enrolling subjects in wasted research may exacerbate mistrust of research among U.S. populations that have borne a disproportionate share of these abuses, namely populations that are poor and populations of color. Government-sponsored research with little prospect of social value is especially sensitive in this larger context of government mistrust and experimentation—particularly if the results are later used for ratcheting in ways that harm funding priorities for these groups.

How much practical benefit is necessary? Most research ethics guidance allows wide latitude in determining that research is beneficial. Value could derive from testing programs to identify immediate benefits in health or welfare, from gaining a better understanding of the causes of problems, or

---

219. Kay Dickersin & Drummond Rennie, *Registering Clinical Trials*, 290 JAMA 516 (2003).

220. Howard Mann, *Research Ethics Committees and Public Dissemination of Clinical Trial Results*, 360 LANCET 406 (2002); Richard S. Saver, *Medical Research and Intangible Harm*, 74 U. CIN. L. REV. 941 (2006); Anna Mastroianni & Jeffrey Kahn, *Swinging on the Pendulum: Shifting Views of Justice in Human Subjects Research*, 31 HASTINGS CTR. REP. 21 (2001).

221. For an introduction to unethical and abusive research practices in the United States, see HARRIET A. WASHINGTON, MEDICAL APARTHEID (2008).

from developing new hypotheses about possible program approaches.[222] Assuming a sound research design and public reporting of findings, evaluations that are neglected by regulators and legislators may nonetheless yield important findings that can be used to shape advocacy efforts or enable variations in programs that are decentralized. But where evaluation designs are unsound, or where results are likely to be distorted or used for ratcheting in ways that disadvantage the research participant populations, the ethical benefits of experimentation are unclear and deserve close attention.

Currently there is little oversight over whether research poses social benefits. Federal regulations of human subjects have delegated responsibility for scrutinizing research protocols to institutional review boards (IRBs). IRBs are generally instructed not to consider the long-term consequences of research during protocol review; the Common Rule in fact directs that "in evaluating risks and benefits, the IRB . . . should not consider possible long-range effects of applying knowledge gained in the research (for example, the possible effects of the research on public policy) as among those risks that fall within the purview of its responsibility."[223] But IRBs do, in fact, often consider social implications of research findings despite this rule.[224] Some have suggested that this is the proper purview of research advisory committees instead.[225] But in the main, it is currently up to individual researchers and funders to consider the potential impacts of their research, including the chance that it will exert little effect on policy, or the chance that it will be used for outsized purposes that harm the constituencies in the research population. In areas that are susceptible to sham EBPM, ensuring rigorous research designs and public availability of research results will enable some of the positive externalities described below, even if the findings are ignored or misused at the governmental level.

---

222. *See* Ezekiel J. Emanuel et al., *What Makes Clinical Research Ethical?*, 283 JAMA 2701 (2000).

223. Criteria for IRB Approval of Research, 45 C.F.R. § 46.111 (2018).

224. Alan Fleishman et al., *Dealing with the Long-Term social Implications of Research*, 11 Am. J. Bioethics 5 (2011); Robert Klitzman, *How IRBs View and Make Decisions About Social Risks*, 8 J. Empirical Res. Hum. Res. Ethics 58 (2013).

225. *Id.*

2. The Government as Experimenter

When the government is acting as experimenter, either by itself or by contracting with researchers, the ethical basis for experimentation may deserve special scrutiny. Compared to other institutions that conduct and sponsor research, the government possesses greater coercive authority, as well as greater structural obligations to promote the well-being of citizens generally, including those who become research subjects. In some research contexts, our faith that the government will carry out its role to safeguard citizens' well-being is the basis for wholesale exemptions from independent ethical review of research designs.

The clearest example of this is an exemption embedded as part of the "Common Rule," the federal regulations providing for institutional review board (IRB) oversight of research with human subjects.[226] In 1983, the Reagan Administration sought to grant a series of Medicaid waivers through § 1115 of the Social Security Act, which allows states a five-year exemption from certain federal regulations in order to implement experimental Medicaid and cash welfare (then AFDC) programming. At the time, the waivers intended to test methods of cost-sharing among the Medicaid population, and were framed in terms of making the programs financially sustainable.[227] The Carter administration had required separate IRB review for experimental Medicaid waivers, but Reagan's HHS quickly reversed this policy. Shortly after the change of administration, HHS implemented changes to the Common Rule with immediate effect,[228] then issued rules specifying that IRB approval would be duplicative and unnecessary for § 1115 waivers. Instead, agency personnel would consider ethics concerns as part of the waiver approval process, which rests on whether a waiver is "likely to advance the purposes" of the Medicaid statute.[229] As the administration noted, "ethical and other research problems raised by

---

226. To What Does this Policy Apply, 45 C.F.R. § 46.101 (2018).

227. Waiver of Requirements as Applied to Medicaid Demonstration Projects Involving Cost-Sharing (Copayments, Deductibles, Coinsurance), 47 Fed. Reg. 9208 (Mar. 4, 1982).

228. *Id.*

229. Exemption of Certain Research and Demonstration Projects from Regulations for Protection of Human Research Subjects, 48 Fed. Reg. 9266 (Mar. 4, 1983); *see also* Lucy A. Williams, *The Abuse of Section 1115 Waivers: Welfare Reform in Search of a Standard*, 12 YALE L. & POL'Y REV. 8 (1994); Sara Rosenbaum, *Mothers and Children Last: The Oregon Medicaid Experiment*, 18 AM. J.L. & MED. 97 (1992).

research in benefits programs will be addressed by the officials who are familiar with the programs."[230] Recent updates to the Common Rule in January 2017 have retained and clarified this exemption from IRB review "in order to make it plain that such research projects on public benefit or service programs [under § 1115 of the Social Security Act] qualify for the exemption."[231]

Separately, a 1992 statutory requirement also specifies that HHS funds may not be used to support research that "poses a danger to the physical, mental, or emotional well-being of a participant without the participant's informed consent,"[232] and agencies considering public benefits experiments must evaluate research according to this standard. In interpreting the language of the Common Rule, the Ninth Circuit's 1994 opinion in *Beno v. Shalala* (along with other courts who have considered the issue) has not challenged the exemption of agency-approved public benefits research from IRB review. Instead, the court noted that the Secretary of HHS was required to "make some determination that a project does not pose unnecessary risks to human subjects" as part of waiver approval, including considering welfare recipients' objections to proposed waiver plans.

This exemption from IRB review is a departure from ordinary practice for research ethics, but reflects the view that government will seek to prioritize the well-being of the populations it serves, including those that it enrolls in experimental programming and evaluations thereof. Here, the risks to subjects may arise not only from those inherent to the program, but also from how decision-makers are likely to use the evidence produced. Where research is so poorly designed as to be inconclusive, or where it is likely to be used disingenuously for ratcheting or distortion, the ethical basis for this research may be suspect, and it is particularly relevant when the government itself is the party experimenting.

E. Positive Externalities

Sham EBPM and broken experimentation have little direct benefit, but they may have silver linings. Here I consider the potential benefits that may arise from even minimal efforts to use and produce evidence, despite the

---

230. 48 Fed. Reg. 9268; *Beno v. Shalala*, 30 F.3d 1057 (1994).

231. Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7195 (Jan 19, 2017).

232. Prohibition on Funding Certain Experiments Involving Human Participants, 42 U.S.C. § 3515b (2018).

fact that evidence may be largely neglected or misused by legislators or regulators making the primary decision.

1. Collateral EBPM

Although the forms of sham EBPM differ, all involve the framing of policy decisions as justified by research evidence. Citing research evidence to justify governmental choices—rather than, say, justifying those choices in terms of pure values or political economy—may in theory promote EBPM by other entities, including more genuine versions. This may happen in several ways.

First, articulating policy choices on empirical grounds—even if those grounds are misconstrued—may have the salutary, inadvertent benefit of signaling to other decision-makers that it is desirable to make policy decisions on the basis of empirical research. This could, of course, lead to the proliferation of sham EBPM practices (a negative externality). A greater prevalence of policies justified as "evidence-based," however, may encourage some observers to engage more fully with research evidence in their own decision-making, including observers who do so effectively and in good faith.

Second, sham EBPM practices may invite political rebuttals that publicize a more faithful depiction of empirical evidence, enabling EBPM elsewhere. For example, an attempt to justify public spending on "Scared Straight" programming for preventing juvenile offending may be couched in terms of anecdotal evidence (perhaps evidence from a hit TV show).[233] But this would mischaracterize the evidence base for Scared Straight programming, which has been shown in multiple randomized studies to exacerbate rather than deter juvenile offending.[234] The continued use of Scared Straight programming on the basis of flimsy evidence has drawn public attention, with broader dissemination of accurate information about research findings.[235] These public correctives may publicize evidence that

---

233. *See, e.g.*, Andy Denhart, *Beyond Scared Straight's Real-Life Controversy*, DAILY BEAST (Jul. 13, 2017) https://www.thedailybeast.com/beyond-scared-straights-real-life-controversy [https://perma.cc/ZP54-SEHB].

234. Petrosino et al., *Scared Straight and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency: A Systematic Review*, 9 CAMP. SYS. REV. 1 (2013).

235. *See, e.g.*, Gary Van Landingham, *Evidence Trumps Anecdote for "Scared Straight" Programs*, PEW (Oct. 14, 2015) https://www.pewtrusts.org/

would not otherwise be brought to light, enabling more evidence-informed decisions elsewhere or at future points in time. Decision-makers who couch bad decisions in evidence-based terms invite a response on the basis of that evidence, which may add motivation for subsequent reform (or perhaps momentum for replacement of the decision-makers).

2. The Hawthorne Effect

Programs may function better simply because they are being observed, regardless of what happens to the evaluation. The marginal improvement in program effectiveness by virtue of being observed is reflective of the Hawthorne effect,[236] also known as observation bias. The first studies of this dynamic were among employees of a telephone manufacturing factory, who produced more under intensive monitoring compared to when they were more loosely supervised; modern studies continue to find support for the proposition that participating in research can improve outcomes.[237] Rationales for the mechanism often include social desirability bias among program participants,[238] but in evaluations of social programs, concern among program staff over ongoing funding may be a more powerful explanation. Those who implement government-funded programs have an incentive to deliver higher-quality programming in the presence of an evaluation, given that the findings may be used to inform funding and implementation decisions. This may improve outcomes for people enrolled in an evaluation, even if the study is poorly conducted and even if the results are never used.

---

en/research-and-analysis/blogs/stateline/2015/10/14/evidence-trumps-anecdote-for-scared-straight-programs [https://perma.cc/9KNT-MGM7]; Denhart, *supra* note 233; *Scared Straight*, WASHINGTON STATE INSTITUTE FOR PUBLIC POLICY (2018) http://www.wsipp.wa.gov/BenefitCost/Program/114 [https://perma.cc/85P9-HZ2V].

236. Rob McCarney et al., *The Hawthorne Effect: A Randomised, Controlled Trial*, 7 BMC RES. METHODOLOGY 30 (2007); Jim McCambridge et al., *Systematic Review of the Hawthorne Effect: New Concepts Are Needed to Study Research Participation Effects*, 67 J. CLIN. EPIDEMIOLOGY 247 (2014).

237. McCambridge et al., *supra* note 236, at 271.

238*. Id.*

3. Clarifying Values

Misuses of evidence may actually have a useful instructive function: they teach us where evidence may not be the right means, or right set of arguments, for deciding a particular policy. Where evidence exists but is neglected, it is a signal that evidence-based reasoning is perceived to be inappropriate for the problem; the problem is one of values, and perhaps cannot be resolved through a technocratic approach of identifying the most effective policy choice.

Even when decisions are exclusively driven by values, however, experimentation may assist current policy development by clarifying the stakes of those choices. Programs may not work as intended, or they may have important externalities that matter and implicate values that come in tension with policy goals. Identifying those externalities is an important benefit of evaluations, and may arise even from evaluations that are of low methodological quality. Say, for instance, a legislator is in favor of criminal penalties for people who expose others to HIV without disclosure, but is also opposed to inequality on the basis of race and gender. Studies of HIV-specific criminal statutes showing disproportionate effects on populations of women and people of color can illuminate the stakes of the policy choice, even if the study isn't the primary basis of the legislative decision.[239] The decision may yet be based entirely on values—namely, the comparative value the legislator places on equity compared to her preference for deterring HIV transmission—but she may not have recognized the values conflict without research results. Like collateral EBPM, this function is more clearly served by good evaluation designs, but even methodologically weak studies may provide relevant evidence to frame the values at stake in policy discussions.

IV. Promoting and Preserving Evidence-Based Efforts

The prior Parts have considered the pathologies that ensue when efforts to generate and use evidence fall so far short of the ideal that they are better considered sham practices. Despite some positive externalities, broken

---

239. Y. Tony Yang & Kristen Underhill, *Rethinking Criminalization of HIV Exposure – Lessons from California's New Legislation*, 378 New Eng. J. Med. 1174 (2018); Amira Hasenbush et al., *HIV Criminalization in California: Penal Implications for People Living with HIV/AIDS*, Williams Inst. (Dec. 2015), https://williamsinstitute.law.ucla.edu/research/health-and-hiv-aids/hiv-criminalization-in-california-penal-implications-for-people-living-with-hivaids/ [https://perma.cc/JJ49-5VDG].

experimentation and sham EBPM may do more harm than good. The pathologies of sham EBPM can undermine scientific legitimacy, chill rigorous research, spend limited governmental funds on ineffective programming (or, in ratcheting, culminate in cutting entire categories of programming that may in fact have social benefits), threaten the balance of benefits and burdens imposed on research subjects, and ultimately result in inefficient uses of public resources with consequences for well-being.

Against this background, I here aim to generate strategies that might promote genuine engagement with EBPM and experiments to evaluate policy decisions. These present both legal and practical means to bolster careful and rigorous uses of evidence and to minimize sham EBPM, broken experimentation, and attendant harms. There is little, in fact, to be done to avoid sham EBPM in areas of highly contested values, but efforts to maximize evaluation quality can (and should) take center stage. Where values are more aligned, steps to promote EBPM through ex ante policy commitments, evaluation mandates, and evaluation transparency hold some promise. The focus herein will be on actions available to government actors, rather than private actors, although private entities may also seek to shape evidence-based policy development.

A. Terms of Engagement: The Limits of EBPM and Experimentation

Where decisions rest on fundamental values that are not shared, EBPM will do little; for these areas of policy choices, the risks of sham EBPM are high. Recall the difference between abortion politics and tornado politics above; although state legislative restrictions on abortion are increasingly couched in the *language* of empirical evidence and women's health, these engage only tenuously with evidence on the health consequences of making abortion more inaccessible.[240] Evidence-based policy-making is unlikely in a space where values are so contested. In recent decades, the U.S. has experienced increased polarization characterized by greater ideological distance between partisans and greater mutual dislike between partisans.[241] New issues have also gained attention from the left and

---

240. CAROL SANGER, ABOUT ABORTION: TERMINATING PREGNANCY IN TWENTY-FIRST-CENTURY AMERICA (2017).

241. Yphtach Lelkes, *Mass Polarization: Manifestations and Measurements*, 80 PUB. OPINION Q. 391 (2016); *see also* Liliana Mason, *Ideologues without Issues: The Polarizing Consequences of Ideological Identities*, 82 PUB. OPINION Q. 866 (2018) (describing antipathy between Democrats and Republicans in terms

resistance from the right in many countries—including environmentalism, access to health care, and a greater focus on rights for women, people of color, and sexual minority groups.[242] In this environment, many issues now verge closer to abortion politics than to tornado politics, characterized by strong and divergent beliefs in what "the evidence" shows.

In a 2011 article describing law as a species of politics, Jeffrey Rachlinski notes why "evidence-based law" is likely to remain an unrealized ideal:

> [B]eliefs are important to people and are resilient . . . . [P]eople who support the death penalty nearly uniformly believe that it deters crime. They also tend to believe that gun ownership makes society safer, that abortion should be illegal . . . and that climate change is not a serious problem . . . . The lack of shared goals means that many studies are essentially irrelevant to underlying legal policy . . . . People interpret social science evidence in ways that are consistent with their beliefs, embracing work that supports them and rejecting work that does not.[243]

Others have commented on the difficulty in reasoning on the basis of empirical evidence where there are fundamental underlying disputes about values.[244] In family law, for example, Claire Huntington notes that the consideration of evidence will be counterproductive without a shared agreement on values; in education law, Eloise Pasachoff notes that the Every Student Succeeds Act (ESSA)—which requires states to base educational reform activities on "evidence"—will do little where there is disagreement on what the purposes of education should be.[245]

Areas of highly contested values are inhospitable to evidence. The problems of bounded rationality of policy-makers are amplified in these settings, and ineluctable boundaries on scientific objectivity, capacity to measure meaningful values, and scientific authority provide valuable

---

of identity, rather than disputes about particular issues). Even where there may be common ground in policy preferences, ingroup and outgroup identities and mutual resentment may complicate compromise).

242. *See, e.g.*, ALEJANDRO MORENO, POLITICAL CLEAVAGES: ISSUES, PARTIES, AND THE CONSOLIDATION OF DEMOCRACY (2019).

243. Jeffery J. Rachlinski, *Evidence-Based Law*, 96 CORNELL L. REV. 901, 921-22 (2011).

244. Eloise Pasachoff, *Two Cheers for Evidence: Law, Research, and Values in Education Policymaking and Beyond*, 117 COLUM. L. REV. 1933 (2017); see also Huntington, *supra* note 171.

245. Pasachoff, *supra* note 244.

materials for motivated political actors. These are prime conditions for evidence distortion, terminal experimentation, and ratcheting, as well as technical breakdowns resulting from motivated interpretations of evidence. Empirical evidence itself is unlikely to change the outcomes of these decisions; such policy choices, particularly in highly visible legislative and regulatory decisions, are made for political reasons and are better explained (and argued) in political economy terms. In the case of ratcheting cuts to after-school programs, for example, "[t]he Trump Administration is not going to be convinced to fund afterschool programs with a data dump of studies showing that the programs are successful. Instead, if the Administration changes its position on whether to fund a given program, it will be due to voters' moral outrage."[246] The Congress that banned further study of Upward Bound programming would not have reallocated funding after another study showing lack of effect; that was a values-driven decision about funding priorities, made on the basis of political rather than evidentiary concerns.

As a practical matter, what should be done to guard against sham EBPM in these areas? There is little hope of salvaging EBPM where values are highly discordant. Legislative mandates to consider and apply evidence, such as ESSA, are likely to culminate in a values-motivated reading of evidence that reinforces legislators' existing positions. But although there is little to do about sham EBPM where values are divergent, it may be more feasible to promote evaluation quality. Where evaluations of policy innovations are done in highly contested areas, the stakes are higher for rigorous designs that minimize bias. This entails independent evaluators, randomization where practicable, statistical power to detect small changes, transparency of evaluation methods, and public release of evaluation findings to enable others besides the primary policy-makers to use the findings. Where evidence is generated on hotly contested policy decisions, particularly evidence from government-mandated or government-sponsored evaluations, ensuring that the study is robust will minimize (but not avoid) sham EBPM. Researchers embarking on these studies should consider the ethical balance of harms and potential benefits, given the possibility for results to be distorted in ways that disadvantage the populations under study. Publicity of findings is critical given that the benefits of such studies are likely to be in their externalities (collateral EBPM), rather than in their primary impact on policy.

In areas more akin to tornado politics, the decision-making environment may be more hospitable to evidence-based decision-making.

---

246. *Id.* at 1969.

EBPM may be more useful for some areas of decision-making than others, even in areas where values diverge. When there is a shared view that a problem exists, evidence can identify the scope and causes thereof (although causes are more controversial than scope). When there is shared agreement on a set of appropriate and politically palatable solutions to a problem, evidence can identify the feasibility and likely effectiveness of solutions in that set. There may also be a divergence in policy views, but shared trust in a research design. Where this occurs, evaluation may work as a tiebreaker or a form of dispute resolution.[247] In these areas, those tasked with making policy choices may be more comfortable engaging in ex ante commitments, mandating robust evaluations, or making ex post disclosures about the evidence base for their decisions.

B. Ex Ante Commitments

In areas where EBPM is feasible, one way to avoid sham versions is to ask decision-makers to make *ex ante* commitments to take action on the basis of evaluation results. That is, legislators or regulators could agree to policies on a trial basis with an evaluation attached. But before the evaluation takes place, they could publicly state their planned future course of action depending on what the results show (including, perhaps, plans to do nothing because they prefer to decide on the basis of values rather than evidence). These statements need not be binding to be useful; nonbinding statements can also call attention to choices that depart from ex ante plans. But a binding pre-commitment of support based on the results of program evaluation would be one mechanism to ensure greater engagement in EBPM, with less interference from political concerns. Ex ante commitments to EBPM may also alleviate some of the problems with bounded rationality of policy-makers by encouraging greater ex ante engagement with research methods and designs.

A range of ex ante commitments may be useful. A simple commitment may be a choice to avoid ratcheting. Before an evaluation begins, legislators may publicly promise that they will not use the results of the evaluation to

---

247. Others have considered the capacity for randomized trials to act as a tiebreaker between adversaries who agree on the rigor of the research design: "When robustness considerations are motivated by an adversarial audience with non-common priors, randomization can be interpreted as a way to let parties with diverging priors agree on a process." Abhijit Banerjee et al., *A Theory of Experimenters* 19 (Nat'l Bureau of Econ. Research Working Paper No. 23867, 2017), https://www.nber.org/papers/w23867 [https://perma.cc/6M4Q-EQ74].

change funding appropriations for the entire policy goal. For example, if Congress decides to earmark federal funds for substance use disorder (SUD) treatment with an evaluation of funded programs, they may publicly promise that they will not use the results of any particular program evaluation to cut funding from SUD treatment. They would, however, retain the ability to reallocate funds to different approaches within that category as the evidence warranted. Although this would not foreclose the allocation of funds within the funding category on the basis of values rather than evidence (consider, for example, a reallocation of funds to increased prosecution of drug crimes instead of medication-assisted treatment), it would avoid the more general problem of defunding entire categories of expenditure. This would alleviate some concerns about chilling research, as well as some of the ethical concerns about doing research that presents the possibility of reducing resources for the populations involved.

Another ex ante commitment is a decision to impose set timelines for renewing specific program approaches (within program categories, which is different from sunsetting an entire category of expenditures), but with the public commitment to review the results of program evaluations at a set time before renewal. This could even include a public declaration of what outcomes would factor into the renewal decision or a stated presumption of renewal (with a specification of what types of evidence would overcome the presumption). Although this would not eliminate sham EBPM (terminal experimentation may be a continued concern, as would evidence distortion if the choices turn out to be more values-based than evidence-based), it would also create the infrastructure and opportunity for genuine EBPM for decisions that are less controversial.

It is also possible to imagine ex ante, pre-evaluation agreements about the types of disclosures that should be made to program participants or their communities. For example, legislators may agree ex ante that if an evaluation finds a program to be ineffective for its stated purpose (e.g., a school voucher program does not produce improvements in educational outcomes), that program should be required to disclose the results of the evaluation to new program participants.[248]

Finally, ex ante commitments could in theory entail promises to support (or to withhold support from) a policy or program depending on evaluation results. This would, essentially, entail presumptive voting on the basis of evidence that is yet to be generated. For example, a provision of the Affordable Care Act allows insurance companies to charge smokers 150%

---

248. This is far softer than de-funding programs, but the effects of disclosure are unpredictable and would need further study.

of the premiums charged to nonsmokers in the individual markets.[249] Congress could in theory remove this flexibility and require equivalent premiums for smokers, given preliminary reports that the rule has led to reduced insurance coverage for smokers but no reduction in smoking behavior itself.[250] If Congress did so, individual legislators may publicly announce that they will support reinstating the law if smoking increases by 30% or more as a result of repeal (i.e., setting a threshold for a decision commitment). Or, in a softer version, they might announce that they intend to change their mind (i.e., setting a presumption), but that they would consider all evidence when making this decision.

This last category of commitment would be a mixed bag. It would avert the problem of free-riding on scientific legitimacy, because it would in fact be a promise to act on the results of the scientific study. It would raise the stakes for methodologically rigorous and adequately powered evaluations. For contentious evaluations, those promises could supply decision-makers with motivation to interfere with evaluations over time. This could also exacerbate problems of authority given to experts, as well as concerns about attentiveness to relevant outcomes that matter to constituencies of interest. These pre-commitments could perhaps spark objections on the theory of delegating legislative authority to private entities (namely, the researchers running the evaluation)—making them more defensible as presumptions rather than enforceable promises about how one will vote. A further problem arises from making decisions on the basis of one or a few evaluations in real time, rather than waiting for long-term results or the aggregation of results from multiple studies.

Individuals in power have little incentive to make pre-commitments tied to evaluation results—the motivation for these promises would need to come from political pressure in favor of evidence-based practice or from the need to make a concession in order to secure other political bargains. But these are intriguing thought experiments for promoting evidence-based decisions. Commitments to avoid ratcheting and to consult evidence at specified intervals may be the most feasible of these proposals, and may go a little way toward alleviating sham EBPM and broken experimentation in practice.

---

249. Patient Protection and Affordable Care Act, 42 U.S.C. § 300gg(a)(1)(A)(iv) (2018).

250. Abigail S. Friedman et al., *Evidence Suggests that the ACA's Tobacco Surcharges Reduced Insurance Take-Up and Did Not Increase Smoking Cessation*, 35 HEALTH AFF. 1176 (2016).

C. Evaluation Mandates

In areas that are more hospitable to EBPM, a salient concern is maximizing evaluation rigor, such that the evidence generated about those programs is of high enough quality to enable genuinely evidence-based choices. Attaching robust and methodologically specific evaluation mandates to program and policy decisions is one way to promote this goal. Under the Spending Clause power, Congress has broad latitude to fund and specify conditions for evaluations of its policy decisions. But to date, evaluation mandates are rare and chronically non-specific; where evaluations are required, such as the evaluations of Medicaid experimental waivers, the approval of specific evaluation protocols is often delegated to an agency,[251] without any parameters set by Congress itself. Imposing more mandates with greater specificity can improve the quality of evidence generated, alleviating broken experimentation, and public scrutiny of this research can mitigate evidence distortion, ratcheting, and terminal experimentation. Improving research designs can also alleviate some of the problems of bounded scientific objectivity and bounded rationality of policy-makers, and greater availability of evidence can support political participation by groups whose interests are affected.

Congress knows how to specify evaluation designs—and, in fact, the legislature has demonstrated its ability to bar specific study designs entirely, as with the DC school voucher study and the Upward Bound experiments. Because these were disputes based on values, not evidence, these evaluations may have done little to influence policy. But where EBPM is more feasible, evaluation mandates can do far more to ensure genuine, rather than sham, experimentation. Mandates can provide methodological guidance (e.g., requiring or encouraging randomized approaches), require the use of independent investigators, require pre-publication of evaluation protocols and public release of results, specify the need for power calculations to ensure adequate sample sizes, require the collection of specific outcomes to ensure attentiveness to a range of relevant outcomes (and to facilitate research synthesis, such as meta-analyses of comparable studies), and fund a mix of quantitative and quantitative study approaches in order to capture unexpected benefits and burdens of the policy or program. Evaluations mandates can also require the timely release of evaluation results, such as before program renewal or sunsets take place.

The evaluation mandate for Medicaid § 1115 waivers provides a partial model. After decades of relative secrecy and weak evaluations, the

---

251. GOV'T ACCOUNTABILITY OFFICE, MEDICAID DEMONSTRATIONS, *supra* note 165.

Affordable Care Act tightened statutory requirements for waiver evaluations. Although there are no specifications for evaluation designs, protocols and evaluation timelines must be made publicly available on the CMS website, along with reports of evaluation results. Evaluations must also report on specific outcome categories, such as insurance coverage, access to care, quality of care, and beneficiary satisfaction with the program.[252] Additional requirements imposed by CMS in 2014 include the use of independent evaluators, and CMS has recently issued advisory recommendations that request rigorous protocols and a mix of qualitative and qualitative designs.[253]

Evaluation mandates—and dedicated funding for evaluation activities—are a feasible strategy to avoid broken experimentation in fields where evidence is likely to be used, either for primary or collateral EBPM. Where the existence, methods, and findings of evaluations are made public, these mandates may also be useful in limiting evidence distortion and ratcheting, and they may mitigate some of the concerns about bounded scientific objectivity.

D. Evaluation Transparency, Ex Ante and Ex Post

Even where evaluations are not mandated, an important defense against broken experimentation is ex ante transparency, including making evaluation protocols publicly available before the evaluation takes place. This has several advantages. It usefully exposes designs to the possibility of public input, which can help to identify secondary outcomes of interest and potentially catch errors before designs are launched. But more importantly, it ties researchers to the mast of the evaluation plan. There are many opportunities for researchers to act on biases when reporting an empirical study—for example, cherry-picking and reporting only findings with favorable results, or subgroup comparisons with results in the predicted directions. Ex ante publication of a study protocol, with intended outcomes and comparisons, is assurance against biased reporting, and it ensures that researchers will implement the protocol and report with fidelity to the original design. Along with evaluation mandates, this transparency can also help to mitigate evidence distortion, terminal experimentation, and ratcheting, and it can provide public information to counteract the bounded rationality of policy-makers and concerns about scientific objectivity.

---

252. Kristen Underhill et al., *Fulfilling States' Duty to Evaluate Medicaid Waivers*, 379 NEW ENG. J. MED. 1985, 1986 (2018).

253. CTR. FOR MEDICAID & MEDICARE, *supra* note 137.

This strategy of pre-publication of research protocols arose in medical research, as a Congressional mandate in 1997. In the Food and Drug Modernization Act that year, Congress directed the NIH to create a registry for clinical trials that would be used as part of FDA approval of experimental drugs for patients with life-threatening conditions. In response, NIH launched clinicaltrials.gov in 2000. By 2005, most medical journals began to require all published clinical trials to have pre-registered with clinicaltrials.gov before submitting manuscripts for publication. The World Health Organization similarly encouraged trial registration and developed a means of searching across multiple trial registries in 2007, the same year in which Congress expanded registration requirements for additional FDA trials, including publishing trial results on the site.[254] Most recently, in July 2018, registration with clinicaltrials.gov or regulations.gov became mandatory for all clinical trials subject to Common Rule research ethics oversight.[255] Although compliance with clinicaltrials.gov reporting requirements is incomplete,[256] and inconsistent reporting still exists,[257] these inconsistencies are comparatively rare; there is widespread agreement that pre-publication is an imperative means of guarding against selective reporting and concealment of trial results.[258]

Although many evaluations of policies and programs are not randomized trials, prior publication of government-sponsored evaluations would help to minimize the problems of broken experimentation. Even for evaluations that are not themselves mandated, it would be straightforward to mandate the pre-publication of all government-sponsored evaluation

254. Food and Drug Administration Amendments Act, Pub. L. No. 110-85, 121 Stat. 823 (2007).

255. *History, Policies, and Law*, CLINICALTRIALS.GOV, https://clinicaltrials.gov/ ct2/about-site/history [https://perma.cc/KWB8-XZGL].

256. Andrew P. Prayle et al., *Compliance with Mandatory Reporting of Clinical Trial Results on ClinicalTrials.gov: Cross Sectional Study*, 344 BMJ, Jan. 3, 2012, at 1; Joseph S. Ross, *Trial Publication After Registration in ClinicalTrials.gov: A Cross-Sectional Analysis*, 6 PLOS MED., Sept. 8, 2009, at 1.

257. Daniel M. Hartung et al., *Reporting Discrepancies Between the ClinicalTrial.gov Results Database and Peer-Reviewed Publication*, 160 ANNALS INTERNAL MED. 477 (2014); Caroline Riveros et al., *Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals*, 10 PLOS MED., Dec. 3, 2013, at 1; Jessica E. Becker et al., *Reporting of Results in ClinicalTrials.gov and High-Impact Journals*, 311 JAMA 1063 (2014).

258. Deborah A. Zarin et al., *Update on Trial Registration 11 Years After the ICMJE Policy Was Established*, 376 NEW ENG. J. MED. 383 (2017).

protocols, timelines, and outcomes on clinicaltrials.gov or an analogous site. Subsequent publication of results in the same place would also help serve the function of releasing findings to the public, facilitating EBPM efforts where possible, or enabling collateral EBPM or momentum for policy change where primary policy choices are hostile to evidence. Mandates for evaluation, whether set by agency rules, legislation, or executive order, could incorporate transparency language that builds on this approach.

E. Teams of Rivals in Evidence Production

The application of evidence is likely to play out in an adversarial process, with competing political priorities and interest groups at stake. But the *production* of evidence on individual research teams is often less controversial, which can compound concerns about scientific objectivity. One strategy that could improve scientific objectivity in the generation of evidence is adversarial collaboration[259] (also described as proponent-skeptic collaboration).[260] This approach requires research teams to include researchers that are ideologically supportive and those that are skeptical of the policy at hand, on the theory that biases at the two ideological poles will make methodological choices as rigorous and objective as possible. An adversarial collaboration model could strengthen rigor at every stage of evaluation, including designing the study, overseeing the protocol, creating an external advisory board, producing research reports. This could contribute to higher-quality research inputs, and it could reduce the possibility of broken experimentation.

---

259. *See, e.g.*, Ian Bateman et al., *Testing Competing Models of Loss Aversion: An Adversarial Collaboration*, 89 J. PUB. ECON. 1561 (2005); Daniel Kahneman, *Experiences of Collaborative Research*, 58 AM. PSYCHOL. 723, 729 (2003); G.P. Latham et al., *Resolving Scientific Disputes by the Joint Design of Crucial Experiments by the Antagonists: Application to the Erez-Latham Dispute Regarding Participation in Goal Setting*, 73 J. APPLIED PSYCHOL. 753 (1988). Dora Matzke et al., *The Effect of Horizontal Eye Movements on Free Recall: A Preregistered Adversarial Collaboration*, 144 J. EXPERIMENTAL PSYCHOL. e1 (2015); Marilyn Schlitz et al., *Of Two Minds: Sceptic-Proponent Collaboration within Parapsychology*, 97 BR. J. PSYCHOL. 313 (2006). *See also* Scott Alexander, *Adversarial Collaboration Contest Results*, SLATE STAR CODEX (Sep. 26, 2018), https://slatestarcodex.com/2018/09/26/adversarial-collaboration-contest-results [https://perma.cc/5MTN-4N7X]. I am especially grateful to participants in the faculty workshop at Northwestern Law for this suggestion.

260. Nuzzo, *supra* note 51, at 185.

F. Judicial Review and Agency Oversight of Evidence-Based Decisions

Judicial review could serve as another means of encouraging more robust engagement with empirical evidence, particularly for agency decisions.[261] Under the Administrative Procedure Act, agency rulemaking and adjudicatory decisions are presumed to be reviewable;[262] courts must set aside agency actions determined to be "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law,"[263] and must also set aside formal adjudicatory decisions that are "unsupported by substantial evidence."[264] Although courts tend to defer to agencies' interpretations of their own statutes (reviewed on an arbitrary and capricious standard)[265] and regulations (reviewed for plain error or inconsistency with the regulation),[266] the availability of judicial review is a means of challenging agency decisions that depart sharply from evidence in the administrative record.

The standard of review for agency rules is by no means that of "ideal" EBPM. The bar for a finding of arbitrariness and capriciousness is fairly high, encompassing actions when the agency "has relied on factors which Congress had not intended it to consider, entirely failed to consider an important aspect of the problem, offered an explanation for its decision that runs counter to the evidence before the agency, or is so implausible that it could not be ascribed to a difference in view or the product of agency expertise."[267] Courts must not substitute their judgment for that of the agencies—there must be "a rational connection between the facts found and

---

261. I am grateful to Matt Spitzer for suggesting the *State Farm* and *Bechtel* cases considered here.

262. APA § 701; *see also* Citizens to Preserve Overton Park v. Volpe, 401 U.S. 402 (1971).

263. APA § 706(2)(A).

264. APA § 706(2)(E).

265. Chevron v. Nat. Res. Defense Council, 467 U.S. 837 (1984). *Chevron* deference is unavailable where an agency either lacks authority delegated by Congress to make rules in the area, or where the agency is not acting pursuant to its delegated authority. United States v. Mead Corp., 533 U.S. 218 (2001).

266. Auer v. Robbins, 519 U.S. 452 (1997); Bowles v. Seminole Rock & Sand Co., 325 U.S. 410 (1945).

267. Motor Vehicle Mfrs. Assn. of U.S., Inc., v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983).

the choice made."[268] But within these contours, courts can reach some of the most egregious instances of ratcheting and evidence distortion.

One such episode was the 1981 decision by the National Highway Traffic Safety Agency (NHTSA) to rescind an agency rule, previously blessed in judicial review, that had required automobile manufactures to equip cars with passive restraints (airbags or passive seatbelts).[269] NHTSA had established the rule in 1977 during the Carter administration, on the assumption that a majority of manufacturers would use airbags, but instead the vast majority of cars were built with seatbelts that could be left unlatched (thus defeating their safety benefits). Under the Reagan administration in 1981, NHTSA concluded that the rule was ineffective and revoked it in its entirety, citing the burdens on the automobile industry.[270] This might be considered an extreme example of ratcheting—because there was evidence that *some* consumers did not use the detachable seatbelts that they had been provided, NHTSA eliminated the mandate for *all* passive restraints. But in concluding that the rule would be ineffective, the agency did not consider two possible alternatives: modifying the rule to require either nondetachable seatbelts or mandatory installation of airbags. When a group of automobile insurers sued for review of the rescission, the Court agreed that it had been arbitrary and capricious. Writing for the majority, Justice Wright explained that the failure to consider the alternatives (mandating nondetachable seatbelts or airbags) was a wholesale abandonment of agency discretion. Where consideration of an obvious alternative was nonexistent, the decision could not have been "the product of reasoned decisionmaking."[271]

Moreover, when NHTSA sought to emphasize its uncertainty about whether consumers would choose to use passive seatbelts, the Court noted that the agency had provided no evidence in support of these doubts: "Recognizing that policymaking in a complex society must account for uncertainty . . . does not imply that it is sufficient for an agency to merely recite the terms 'substantial uncertainty' as a justification for its actions. The

---

268. *Id.*, at 43.

269. *Id.*

270. Justice Rehnquist's dissent notes that the decision was likely "related to the election of a new President of a different political party" with a different "philosophy" on the issue. For Rehnquist, political change was "a perfectly reasonable basis for an executive agency's reappraisal of the costs and benefits of its programs and regulations," so long as any change was "within the boundaries established by Congress." *Id.*, at 59.

271. *Id.*, at 52.

agency must explain the evidence which is available."[272] The court notes, too, that "the empirical evidence on the record . . . reveals more than a doubling of the usage rate experienced with manual belts."[273] Without engaging with this evidence, NHTSA had departed so sharply from an "evidence-based" decision that it met the high bar for arbitrary and capricious action.

The *State Farm* decision leverages the arbitrary and capricious standard of review in two ways, both of which can bring agency actions closer to evidence-based practice. First, the court requires the agency to consider obvious alternatives (such that the agency is in fact exercising discretion), and second, the court identifies decisions untethered to the evidentiary record as arbitrary and capricious actions that exceed agency authority. These tools are limited, of course, to the evidentiary record before the court—where evidence is not introduced in the administrative record, such as through notice-and-comment rulemaking, it would be more difficult to courts to identify its absence, and to therefore fault agencies for lack of consideration. But where evidence is available, judicial review may provide some check on ratcheting. Judicial review may also police evidence distortion, although to a lesser extent. Courts are perhaps poorly suited to instruct agencies on how to weight the evidence that they review, and they are at a disadvantage in instructing agencies about evidence missing from the administrative record. But where there is a great disparity between the evidence in the record and the decision reached, the ability of courts to reverse actions that "run counter to the evidence" can be a tool (however weak) against evidence distortion.[274]

Courts have less capacity to rein in the shortcomings of EBPM efforts by legislatures. Assuming legislation does not implicate a fundamental right or a suspect class, statutes are reviewed only for rational basis, under which "legislation is presumed to be valid and will be sustained if the classification

---

272. *Id.*

273. *Id.*, at 53.

274. Judicial review is likely to be less effective against terminal experimentation. Terminal experimentation would likely be borne out in challenges based on a failure to issue rules, which may only be actionable if it frustrates the operation of a federal statute. The decision to refrain from rulemaking, however (such as by rejecting a party's petition to issue rules), is "'at the high end of the range' of levels of deference given to agency action," actionable only when the agency has failed to "adequate[ly] explain the facts and policy concerns it relied on and . . . those facts have some basis in the record." *See* Preminger v. Sec'y of Veterans Affairs, 632 F.3d 1345 (Fed. Cir. 2011).

drawn by the statute is rationally related to a legitimate state interest."[275] Congress has a wide berth for error on this standard, by which it is presumed that "even improvident decisions will eventually be rectified by the democratic processes."[276] Legislation that burdens fundamental rights or a suspect class is reviewed based on whether it is "narrowly tailored" to a compelling state interest, but courts do not scrutinize whether Congress has chosen the option with the strongest evidence base. Without more tools, the pathologies of evidence distortion, ratcheting, and terminal experimentation are largely out of reach with respect to Congressional decisions.

Beyond judicial review, agency watchdogs may also serve as checks on irresponsible uses of evidence. The Government Accountability Office (which provides feedback to Congress on federal expenditures) and the Congressional Budget Office (which provides analyses of economic issues including projections of the impacts of proposed legislation) might promote more rigorous uses of empirical evidence by assessing the likely impact *of* government-funded evaluation research, in light of research designs, transparency of findings, and political environment for using research findings. CBO already plays an important role in quantifying the likely impacts of proposed legislation, which is central to EBPM efforts. The Office of the Inspector General in each agency (which oversees agencies for fraud and waste) might also broaden its consideration of possible evaluation funds as "waste" when they are unlikely to lead to usable evidence.

CONCLUSION

Evidence-based policy is on the rise, and legislation and agency regulation have been no exception to calls for greater uptake of research in practice. Indeed, modern interests in moneyball and EBPM are part of a long history of efforts to promote research-based decisions in government. Where research is inconclusive, as in the case of the NRC death penalty report, many are left feeling rudderless. But there are many reasons to be skeptical of the capacity of EBPM in governmental decision-making. EBPM is itself bounded by limits on rationality, the capacity of science, the objectivity of science, and the authority we wish to give technocrats. Highly values-driven decisions will be resistant to evidence-based reasoning and are more likely to result in "sham" evidence-based practices, such as evidence distortion, terminal experimentation, and ratcheting actions to

---

275. Cleburne v. Cleburne Living Ctr., 473 U.S. 432, 440 (1985).

276*. Id*.

sanction entire funding priorities rather than individual programs. Broken experimentation is also common, with evaluations of government programming and policies missing innumerable opportunities to provide rigorous evidence.

This Article has considered the effects of broken experimentation and sham EBPM, including potential harms to welfare, scientific legitimacy, research subjects, and the production of research evidence, as well as potential downstream benefits through collateral EBPM where evaluations are reliable. Given the balance of harms and downstream benefits, I have proposed a set of strategies to promote genuine EBPM (where possible) and rigorous evaluations of policy choices. Some are more feasible, such as more specific evaluation mandates, adversarial collaborations, or the pre-publication of government-sponsored program and policy evaluations on a repository such as clinicaltrials.gov. Some are less likely, such as pre-commitments to avoid ratcheting, to require disclosures in governmental programs, or to support policy choices on the basis of impending evaluations. Regardless, the current moment of interest in EBPM is generally worth encouraging, whether by these means or other efforts to promote the uptake of research evidence. Assuming that legislators and agency personnel in fact wish to promote the good of the people, genuine engagement with empirical evidence promises greater efficiency in reaching those goals. These strategies may help avoid the pitfalls of poor experimentation and sham EBPM, and where evidence-based practice is feasible, they may help put research-based choices on more solid ground.