Antifragile Policymaking: A Strategy for Institutional Response to the Social Science Reproducibility Crisis

JOHN S. EHRETT*

* Law Clerk, U.S. Court of Appeals for the Fifth Circuit; Yale Law School, J.D. 2017.

I. THE MAINSTREAMING OF BEHAVIORAL SCIENCE IN LAW

A. Introduction

A specter is haunting policymakers—the specter of unreliable science.

In recent years, one of the most prominent institutional expressions of evidence-based policymaking has been the incorporation of insights derived from the behavioral sciences into law and policy. These insights have often shown great promise, offering the potential for optimized delivery of governmental programs, better insights into human thought processes, and so forth. A burgeoning field of scientific research, however, suggests that many of the underlying studies—including studies bearing on topics of critical national importance—may not report meaningfully generalizable conclusions about social populations writ large. This problem has been labeled a "reproducibility crisis," and this Essay aims to provide an initial evaluation of the downstream consequences of this crisis, as it pertains to behavioral science, for lawmakers and policymakers.[1]

The analysis proceeds in four parts: first, consideration of the ways in which law and policy have moved toward an increasing dependence on social-science findings; second, an outline of the current

---

1. Current reproducibility controversies extend across scientific fields. *See, e.g.*, Monya Baker, *1,500 Scientists Lift the Lid on Reproducibility*, NATURE, http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970 (last updated July 28, 2016). Notwithstanding the broader extent of this problem, this Essay focuses its attention primarily on behavioral science.

The reasons for this orientation are threefold. First, the federal government's efforts to incorporate behavioral science into agency decision-making processes are fairly new, which makes this area particularly ripe for reform. *See infra* note 11 and accompanying text. Second, the use of behavioral science is less broadly entrenched across the legal process than reliance on other, older disciplines. The existence of ongoing debates over the admissibility of testimony from such newer fields testifies to this reality. *See generally, e.g.*, Thomas D. Albright, *Why Eyewitnesses Fail*, 114 PROC. NAT'L ACAD. SCI. 7758 (2017) (discussing the current state of science regarding eyewitness testimony); Dara Loren Steele, Note, *Expert Testimony: Seeking an Appropriate Admissibility Standard for Behavioral Science in Child Sexual Abuse Prosecutions*, 48 DUKE L.J. 933 (1999) (exploring how behavioral science might be employed in the courtroom in specific criminal cases). Third, the problem of reproducibility appears to be most severe—broadly speaking, that is—in the domains this Essay explores. *See infra* note 78.

methodological crisis in these fields; third, explanation of the reasons this crisis may prove so difficult to address from the perspective of governmental organs; and last, the ways in which these organs might move toward a norm of *antifragile policymaking*.

The term "antifragile," as used here, comes from Nassim Nicholas Taleb, who defines "antifragile" as those things that "benefit from shocks" and become stronger when exposed to "volatility, randomness, disorder, and stressors."[2] Effective science-based laws and policies should be antifragile—constructed in such a way as to accommodate changes and become stronger over time as new findings emerge.

If scientific findings are to be increasingly mainstreamed into legal and political institutions, this philosophical approach—antifragile policymaking—should play an important role. Unlike certain value commitments that inform law and policy, science is ideally and necessarily dynamic, and science-based policies must be structured to handle that dynamism. The present failure of agencies (as well as courts and legislators) to design antifragile science-based policymaking structures has subjected institutions to a risk of serious disruption.

Notably, this Essay's call for an adjustment to institutional approaches in no way entails a broad epistemological skepticism about the insights of behavioral science—or scientific evidence—writ large. By contrast, the problem identified here is the fact that current legal structures are simply ill equipped to deal with an upstream methodological crisis. When insights from non-legal disciplines are wedded to the coercive power of law, it is imperative that there be a way to separate the scientific wheat from the chaff. Simply put, evidence-based policies should correspond to reality and the uncritical incorporation of controversial scientific research into law and policy risks thwarting that end.

---

2. NASSIM NICHOLAS TALEB, ANTIFRAGILE: THINGS THAT GAIN FROM DISORDER 3 (2012); *see also* J.B. Ruhl, *Managing Systemic Risk in Legal Systems*, 89 IND. L.J. 559, 587 (2014) ("When the legal system has not succeeded in avoiding a failure in another social system or in the legal system itself . . . . we design more fail-safe strategies to patch up the problems the previous set did not adequately manage.").

B. Translating Behavioral Science into Law and Policy

The relationship between behavioral science and law has deepened over time,[3] but ties have grown particularly strong in recent years. In their 2006 book *Nudge*, social scientists Cass Sunstein and Richard Thaler advanced a vision of benign paternalism in which apparently small-scale policy adjustments can produce large-scale effects across a given public.[4] Construed in the broadest sense, "nudges" of the sort envisioned by Sunstein and Thaler play on unconsciously held preferences to steer individual and group decisions in socially optimal ways. "If we value democratic self-government," Sunstein has argued, "we will be inclined to support nudges and choice architecture that can claim a democratic pedigree and that promote democratic goals."[5]

As an example of one such "nudge," Colin Camerer and others have given the example of *defaults*—frameworks for decision-making that subtly steer individuals' preferences in socially desirable directions.[6] Defaults are based on the widespread belief that when faced with a choice, individuals are less likely to choose an option that departs from the default setting.[7] Consider the example of a default rule aimed to encourage individuals' participation in a retirement savings program: if the choice to participate is initially framed

---

3. The incorporation of behavioral science into legal structures has a long-standing pedigree. *See, e.g.*, Angela L. Sharpe, *Working with the Federal Government*, *in* HANDBOOK ON COMMUNICATING AND DISSEMINATING BEHAVIORAL SCIENCE 251, 252 (Melissa K. Welch-Ross & Lauren G. Fasig eds., 2007) (exploring this history).

4. RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 5–6 (2009) (defining a "nudge" as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid.").

5. Cass R. Sunstein, *The Ethics of Nudging*, 32 YALE J. ON REG. 413, 415 (2015).

6. Colin Camerer et al., *Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism,"* 151 U. PA. L. REV. 1211, 1224–25 (2003).

7. *See* Eric J. Johnson et al., *Defaults, Framing, and Privacy: Why Opting In-Opting Out*, 13 MARKETING LETTERS 5, 6–8 (2002).

as an "opt-in" decision, fewer individuals will choose to participate in the given retirement plan than if the default is an "opt-out" decision that requires individuals to explicitly register their refusal to participate. The end result is a net win for policymakers aiming to increase participation in the savings program, with the added benefit that no direct coercion is required.[8] The program works by sustaining an illusion of unmanipulated choice.[9] In light of such research, Sunstein and Thaler's central justification for "nudges" is their belief that because a great deal of decision-making apparently occurs on a subconscious level, the introduction of subtle environmental or social cues can trigger desirable shifts in expressed preferences or behaviors.[10]

Similar ideas have increasingly gained traction across the policymaking world. Perhaps the crowning moment of the behavioral sciences' assimilation into law and policy was President Barack

---

8. "Nudges" in policymaking are not altogether uncontroversial. *See* CASS R. SUNSTEIN, THE ETHICS OF INFLUENCE: GOVERNMENT IN THE AGE OF BEHAVIORAL SCIENCE 34 (2016) ("Efforts to target, or to benefit from, behavioral biases tend to be more controversial on ethical grounds than efforts to appeal to deliberative capacities. The reason is that the former may appear to be more manipulative and less respectful of people's capacity for agency."). At times, opposition to "nudging" manifests as conspiracy theorizing. *See, e.g.*, Alex Newman, *Obama Decree Unleashes "Behavioral Science" Squad to "Nudge" You*, NEW AM. (Sept. 18, 2015), https://www.thenewamerican.com/usnews/constitution/item/21609-obama-decree-unleashes-behavioral-science-squad-to-nudge-you (characterizing the new Social and Behavioral Sciences Team as a vanguard of "federal mind manipulation" that has been "charged with applying psychological research to manipulate the beliefs and behaviors of Americans under various pretexts, and guide the vast federal bureaucracy as it hones its abilities to manipulate the public"). While recognizing the importance and salience of critical discussions in this area, this Essay takes no position on the normativity of "nudge"-type interventions.

9. *See* Jeremy Waldron, *It's All for Your Own Good*, N.Y. REV. BOOKS (Oct. 9, 2014), https://www.nybooks.com/articles/2014/10/09/cass-sunstein-its-all-your-own-good/ ("The result would be a sort of soft paternalism: paternalism without the constraint; a nudge rather than a shove; doing for people what they would do for themselves if they had more time or greater ability to pick out the better choice.").

10. *See* THALER & SUNSTEIN, *supra* note 4, at 3–4. For additional examples, see Pam Belluck, *Reinvent Wheel? Blue Room. Defusing a Bomb? Red Room.*, N.Y. TIMES (Feb. 5, 2009), http://www.nytimes.com/2009/02/06/science/06color.html?_r=0; Benjamin Wallace-Wells, *Cass Sunstein Wants to Nudge Us*, N.Y. TIMES MAG. (May 13, 2010), https://www.nytimes.com/2010/05/16/magazine/16Sunstein-t.html.

Obama's 2015 executive order calling for agencies to employ insights from behavioral science in "encourag[ing] or mak[ing] it easier for Americans to take specific actions."[11] This order, in turn, gave rise to the Social and Behavioral Sciences Team ("SBST"), a body specifically tasked with applying the insights of behavioral science to federal administrative operations.[12] So far, there is no evidence that President Donald Trump has rescinded this particular order. And as written, the text of the order suggests that analysis of behavioral science may become a permanent feature of agency decision-making.[13]

While questions over the relationship between empirical science and law are certainly not new, what *is* novel are these recent

---

11. Exec. Order No. 13,707, 80 Fed. Reg. 56,365 (Sept. 15, 2015) [hereinafter *Executive Order*].

12. *See About SBST*, NAT'L SCI. & TECH. COUNCIL (Jan. 20, 2017), https://sbst.gov/#work ("Accessing and using programs should be intuitive. Information and choices among program options should be clear. Forms should be simple and easy to complete. Behavioral science provides us with tools for designing the kind of government Americans deserve.").

Shortly after the SBST's establishment, one SBST project set out to successfully enroll those Affordable Care Act marketplace participants who had stopped signing up partway through the process. To achieve this, "officials sent out eight versions of letters to applicants, one of which included a photo of the sender to 'personalize the message'" and "[t]hat letter boosted enrollment by 13.2 percent." Jonathan Miller, *Health Insurer Law Test Changes Behavior Policy*, CQ ROLL CALL INS. BRIEFING, Sept. 21, 2015. This led to "roughly 4,930 new enrollments and $1.3 million in savings in just the first month after the emails were sent." April Lea Pope, *To Behave or Not to Behave: How Behavioral Science Can Inform Policy and the Law*, 59 ADVOC. 41, 42 (2016).

In a potentially more controversial move, however, the SBST and the Consumer Financial Protection Bureau have moved to adjust "[t]he 'reasonable consumer' paradigm introduced during the Reagan administration and advocated by the Federal Trade Commission (FTC) . . . into a more behavioral consumer concept." Philipp Hacker, *More Behavioral vs. More Economic Approach: Explaining the Behavioral Divide Between the United States and the European Union*, 39 HASTINGS INT'L & COMP. L. REV. 355, 367 (2016). By doing so, the SBST is using findings from behavioral science to adjust the decision-making standards and concepts by which agencies will be evaluated in judicial review proceedings.

13. *See Executive Order*, *supra* note 11 ("[A]gencies shall consider how the timing, frequency, presentation, and labeling of benefits, taxes, subsidies, and other incentives can more effectively and efficiently promote those actions, as appropriate. Particular attention should be paid to opportunities to use nonfinancial incentives.").

pushes to formally incorporate behavioral science insights into lawmaking and policymaking at the highest federal levels. And as one might expect, an important assumption underlies this push to build "nudges" into policymaking and normalize behavioral science as a basis for government action: the assumption that evidence provided by behavioral science is, generally speaking, just as reliable as evidence gleaned from the physical sciences. This assumption, however, has increasingly been called into question by recent analyses.

II. THE ARCHITECTURE OF AN INTERDISCIPLINARY CRISIS

A. The Reproducibility Crisis

The processes through which scientific research advances—application of the scientific method, analysis, peer review, and others—are predicated on an underlying belief in reproducibility.[14] For example, if a particular default setting for retirement savings plans *really does* affect individuals' decision-making, that effect should show up repeatedly when experiments in default setting are conducted within the same parameters as the original study. Standardized evaluative criteria have been developed to help ascertain whether random chance produced an observed effect or whether a meaningful causal relationship actually does exist between two experimental variables.[15] In some contexts, follow-up studies and meta-analyses have continued to demonstrate the reality of an initially observed effect. Consider the example of loss aversion—the tendency of individuals to prioritize retention of a given item over acquisition of the same item, even when the *ultimate* outcomes are materially equivalent.[16]

14. *See, e.g.*, Arturo Casadevall & Ferric C. Fang, Editorial, *Reproducible Science*, 78 INFECTION & IMMUNITY 4972, 4972 (2010) ("There may be no more important issue for authors and reviewers than the question of reproducibility, a bedrock principle in the conduct and validation of experimental science.").

15. *See* Tukur Dahiru, P-*Value, A True Test of Statistical Significance? A Cautionary Note*, 6 ANNALS IBADAN POSTGRADUATE MED. 21 (2008) (chronicling and critiquing the historical evolution of these norms).

16. Daniel Kahneman & Amos Tversky, *Choices, Values, and Frames*, 39 AM. PSYCHOLOGIST 341 (1984) (introducing this concept).

Numerous meta-analyses of the loss aversion impulse have offered strong support for the view that this phenomenon indeed exists.[17]

Elsewhere, however, the picture of scientific progress is not so clear. A growing body of evidence suggests that much contemporary research in the social and behavioral sciences cannot be replicated when the studies in question are conducted with new participants.[18] Just to name one example, a widely publicized, recent campaign to test the reproducibility of influential studies resulted in the disconcerting finding that "14 of 55 (25%) of social psychology effects replicated by the $P < 0.05$ criterion, whereas 21 of 42 (50%) of cognitive psychology effects did so. Simultaneously, all journals and disciplines showed substantial . . . declines in effect size in the replications compared with the original studies."[19] In other words, within the sample under consideration, half of all cognitive psychology effects and *three-quarters* of social psychology effects could not be fully replicated. The implications of this finding are obviously troubling on their face, particularly for policymakers heavily invested in evidence-based decision-making in regard to behavioral science.

Why might there be such a dramatic discontinuity between initial published results and subsequent reanalysis? While a fully comprehensive treatment of the sociological factors undergirding this challenge is far beyond the scope of this Essay, at least four major aspects of the problem warrant explicit consideration: publication pressures, "*p*-hacking," the non-publication of null results, and the systemic mischaracterization of scientific findings in popular media.

---

17. *See, e.g.*, Nico Neumann & Ulf Böckenholt, *A Meta-Analysis of Loss Aversion in Product Choice*, 90 J. RETAILING 182 (2014) (aggregating research in this area).

18. Wholly independent of those specific challenges identified here, behavioral science already faces the difficulties associated with isolating variables within complex patterns of human behavior. *See, e.g.*, Jerry Adler, *The Reformation: Can Social Scientists Save Themselves?*, PAC. STANDARD (Apr. 28, 2014), https://psmag.com/the-reformation-can-social-scientists-save-themselves-8c2f834715a7#.hgdpezlfy ("Subjects sometimes figure out what's going on and correct for it. Differences in setting, or the selection of subjects, can confound results. Which in a sense just deepens the epistemological quagmire."); *see also* Baker, *supra* note 1 (evaluating the extent of this problem).

19. Open Sci. Collaboration, *Estimating the Reproducibility of Psychological Science*, 349 SCI. 4716-1, 4716-5 (2015).

1. Publication Pressures

University-based researchers face strong institutional incentives—the "publish or perish" dynamic, for one—that reward rapid publication of potentially influential findings. Accordingly, researchers have little incentive to conduct experiments that risk disconfirming particularly intriguing results.[20] As two scholars have written, "[i]nstitutions may . . . serve the goal of improving research credibility and efficiency if they adopt appointment and promotion standards that, instead of relying on publication in top tier journals as a surrogate for quality, recognize[] the importance of reproducible research findings rather than flashy, unsubstantiated reports."[21]

2. *P*-Hacking

In statistical research, *p*-values are probabilities used to identify whether a given correlation between two survey variables is the result of random chance (a "null" result) or reflects a real relationship.[22] *P*-hacking is the practice of adjusting, *ex post*, the parameters of a given research study to identify *some* statistically significant correlation between two variables.[23] *P*-hacking inverts the scientific

---

20. These potential risks do not include the possibility of actual malicious behavior resulting from such pressures. *See, e.g.*, Benedict Carey, *Fraud Case Seen as a Red Flag for Psychology Research*, N.Y. TIMES (Nov. 2, 2011), http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html ("A well-known psychologist in the Netherlands whose work has been published widely in professional journals falsified data and made up entire experiments, an investigating committee has found.").

21. C. Glenn Begley & John P.A. Ioannidis, *Reproducibility in Science: Improving the Standard for Basic and Preclinical Research*, 116 CIRCULATION RES. 116, 124 (2015).

22. *See* Christie Aschwanden, *Not Even Scientists Can Easily Explain P-Values*, FIVETHIRTYEIGHT (Nov. 24, 2015, 12:12 PM), http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/ ("Imagine . . . that you have a coin that you suspect is weighted toward heads. (Your null hypothesis is then that the coin is fair.) You flip it 100 times and get more heads than tails. The p-value won't tell you whether the coin is fair, but it will tell you the probability that you'd get at least as many heads as you did if the coin was fair. That's it—nothing more.").

23. *See* Megan L. Head et al., *The Extent and Consequences of* P-*Hacking in Science*, PLOS: BIOLOGY, Mar. 2015, at 1.

method: in lieu of confirming or disconfirming a research hypothesis formulated *ex ante*, it relies on reverse-engineering a data set to identify correlations that may well prove unreliable.

Given the tendency of such *p*-hacking to generate neutered or deceptive research findings, the American Statistical Association has officially stated that:

> *P*-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "*p*-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided.[24]

3. Non-Publication of Null Results

In light of the institutional incentives identified above, researchers not explicitly engaging in *p*-hacking may simply elect not to publish study results that disconfirm their initial hypotheses. This skews the academic literature: novel correlations and findings are highlighted, while less "interesting" results go unseen.[25]

---

24. Ronald L. Wasserstein & Nicole A. Lazar, Editorial, *The ASA's Statement on* P-*Values: Context, Process, and Purpose*, 70 AM. STATISTICIAN 129, 131–32 (2016).

25. *See* Natalie Matosin et al., Editorial, *Negativity Towards Negative Results: A Discussion of the Disconnect Between Scientific Worth and Scientific Culture*, 7 DISEASE MODELS & MECHANISMS 171, 171 (2014) ("Because scientists are involuntarily finding themselves engaged in competition for positions and funding, many are choosing not to proceed with their non-significant findings (those that support the null hypothesis) that yield less scientific interest and fewer citations. Consequently, the amount of non-significant data reported is progressively declining . . . ."); *see also* Daniele Fanelli, *Negative Results Are Disappearing from Most Disciplines and Countries*, 90 SCIENTOMETRICS 891 (2012) (providing an empirical basis for the foregoing assessment).

4. Misleading Headlines and Unreasonable Expectations

Mass-media culture tends to distort popular perceptions about what scientific research has or has not demonstrated; the exaggerated claims described in such misleading journalism are per se irreproducible, which risks undermining public faith in scientific research writ large.[26]

A great deal has already been written in scientific literature about the dimensions of this reproducibility crisis confronting modern researchers.[27] The implications of this ongoing situation for lawmakers and policymakers, however, have not been fully explored. And the stakes involved in this problem are quite high: consider the highly salient, politically charged issue of implicit racial bias.

B. Case Study: Implicit Bias

Following several controversial, highly publicized police shootings of unarmed black men, implicit bias—unconscious negative psychological associations harbored about members of a given race or other minority group, or based on other traits like class or gender[28]—has become an increasingly salient social and political issue.[29] The foundation for scientific analyses of implicit racial bias

---

26. *See, e.g.*, Robert Gebelhoff, Opinion, *The Media Is Ruining Science*, WASH. POST (Aug. 17, 2016), https://www.washingtonpost.com/news/in-theory/wp/2016/08/17/the-media-is-ruining-science/?utm_term=.a068d46a2080 ("[M]edia agents for research institutions have become adept at turning complicated scientific jargon into compelling press releases—usually at the expense of accuracy. Reporters crop down those releases even further, stretching, exaggerating and torturing academic papers until their original meaning of the study has been completely lost.").

27. *See supra* notes 18–19 and accompanying text.

28. *See* CHERYL STAATS ET AL., STATE OF THE SCIENCE: IMPLICIT BIAS REVIEW 2015, at 4 (2015), http://kirwaninstitute.osu.edu/wp-content/uploads/2015/05/2015-kirwan-implicit-bias.pdf ("Implicit biases are activated involuntarily and beyond our awareness or intentional control. . . . While . . . implicit associations may form as a result of exposure to persistent stereotypes, implicit bias goes beyond stereotyping to include favorable or unfavorable evaluations toward groups of people.") (emphasis omitted).

29. *See, e.g.*, Tom James, *Can Cops Unlearn Their Unconscious Biases?*, THE ATLANTIC (Dec. 23, 2017), https://www.theatlantic.com/politics/archive/2017/12/implicit-bias-training-salt-lake/548996/.

has often been the Implicit Association Test ("IAT"), a Harvard-promoted assessment tool that tracks participants' association of "positive" and "negative" concepts with racial and other group-based descriptors.[30] This tool might detect whether a concept like "white" is more reflexively associated with concepts like "pleasant"—"which might be expected for White subjects raised in a culture imbued with pervasive residues of a history of anti-Black discrimination."[31] IAT-centric research has typically identified such reflexive associations.[32]

Given such a finding, implicit bias is frequently invoked as a powerful driver of persistent racial inequality across diverse societal contexts.[33] Most significantly for policymakers concerned about equitable administration of justice, implicit bias has been connected by extensive academic literature to racial disparities in criminal justice outcomes.[34] The theory intuitively "works": thanks to social conditioning, individuals suffer from persistent tendencies toward uncon-

30. *See* Olivia Goldhill, *The World Is Relying on a Flawed Psychological Test to Fight Racism*, QUARTZ (Dec. 3, 2017), https://qz.com/1144504/the-world-is-relying-on-a-flawed-psychological-test-to-fight-racism/ (noting the spread of the IAT "from Yale's freshmen to millions of people worldwide").

31. Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464, 1465 (1998); *see also* Anthony G. Greenwald et al., *Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm*, 85 J. PERSONALITY & SOC. PSYCHOL. 197 (2003) (further contextualizing the IAT).

32. *See, e.g.*, Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CAL. L. REV. 945, 953 (2006).

33. *See, e.g.*, KIRWAN INST., UNDERSTANDING IMPLICIT BIAS 2 (2012), http://kirwaninstitute.osu.edu/docs/implicit-bias_5-24-12.pdf ("Implicit bias is a mental process that stimulates negative attitudes about people who are not members of one's own 'in group.' Implicit racial bias leads to discrimination against people who are not members of one's own racial group.").

34. While the existing literature on this topic is enormous, several recent works serve as excellent exemplars of this research area. *See, e.g.*, Cedric L. Alexander, *Community Policing as a Counter to Bias in Policing: A Personal Perspective*, 126 YALE L.J.F. 381 (2017); Mark W. Bennett, *The Implicit Racial Bias in Sentencing: The Next Frontier*, 126 YALE L.J.F. 391 (2017); Justin D. Levinson & Robert J. Smith, *Systemic Implicit Bias*, 126 YALE L.J.F. 406 (2017); L. Song Richardson, *Systemic Triage: Implicit Racial Bias in the Criminal Courtroom,* 126 YALE L.J. 862 (2017) (reviewing NICOLE GONZALEZ VAN CLEVE, CROOK COUNTY: RACISM AND INJUSTICE IN AMERICA'S LARGEST CRIMINAL COURT (2016)).

scious prejudice.[35] Thus, in debates about the existence and character of institutional racism, the persistence of implicit bias is often invoked.[36]

The functional appeal of this explanatory framework for policymakers is readily apparent. If social conditioning is responsible for the existence of latent bias, adjustments to the social conditioning process can help push back against racism-producing dynamics. Moreover, implicit bias theory subtly absolves individual actors of guilt. Racialized disparate outcomes can be blamed on an abstracted cultural superstructure, rather than on individuals' consciously held prejudices.

But the usefulness of the "implicit bias" concept as an explanatory paradigm for persistent racial discrimination may prove short-lived; new meta-analyses aggregating implicit bias research have failed to systematically reproduce the effect so often cited by policymakers and science writers.[37] In a groundbreaking recent paper—the

35. *See* Dana Lee Marks, *Who, Me? Am I Guilty of Implicit Bias?*, AM. BAR ASS'N (Nov. 1, 2015), https://www.americanbar.org/groups/judicial/publications/judges_journal/2015/fall/who_me_am_i_guilty_of_implicit_bias/ ("Explicit prejudice is obvious, easily identified, and hopefully rare in judicial settings. . . . In contrast, because implicit bias is automatic and functions below our conscious awareness, it is far more pernicious and dangerous."); Chris Mooney & Indre Viskontas, *The Science of Your Racist Brain*, MOTHER JONES (May 9, 2014, 10:00 AM), http://www.motherjones.com/environment/2014/05/inquiring-minds-david-amodio-your-brain-on-racism ("[O]ur brains have evolved to see patterns in things that are complex, and to categorize the world in order to simplify it. Thus, when we encounter another person, our brains rapidly and subconsciously try to figure out if he or she is friend or foe: in-group or out-group.").

36. *See, e.g.*, Megan Mitchell, Implicit Bias, Colorblindness and Institutional Racism 36 (2014) (unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill), https://cdr.lib.unc.edu/indexablecontent/uuid:8bcbd85c-c7b1-4bf4-85a2-ab6f8f795b02.

37. The problem is not that *no* studies have found a correlation between implicit bias and discrimination but that these studies' effects cannot be generally replicated. *See* Frederick L. Oswald et al., *Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies*, 105 J. PERSONALITY & SOC. PSYCHOL. 171, 188 (2013) ("[I]ndividual studies [found] statistically significant correlations between IAT scores and some criterion measures of discrimination and . . . that IATs had greater predictive validity than explicit measures of bias when predicting discrimination against African Americans and other minorities.") (citations omitted). *But see id.* ("[T]he IAT provides little insight into who will discriminate against whom, and provides no more insight than explicit measures of bias. The

"first large-scale quantitative synthesis of research on change in implicit bias"[38]—a multi-university team of researchers found that "changes in implicit bias did not mediate changes in explicit bias and behavior."[39] In other words, "there is very little evidence that changes in implicit bias have anything to do with changes in a person's behavior."[40]

The theory that implicit bias is a predictor of discriminatory behavior has undergirded an enormous amount of scholarship, policymaking, and critical commentary.[41] A contrary finding of this magnitude—a *dramatically counterintuitive* finding, given the ubiquity of the "implicit bias" theory across legal and policy literature—should logically have colossal ripple effects across the literature on the psychology of discrimination.[42] If the answer to the a priori question—is this belief founded on fact?—changes, this entire institutional edifice is called into question[43]: if policymakers really intend to

---

IAT is an innovative contribution to the multidecade quest for subtle indicators of prejudice, but the results of the present meta-analysis indicate that social psychology's long search for an unobtrusive measure of prejudice that reliably predicts discrimination must continue.") (citations omitted).

38. Patrick Forscher et al., *A Meta-Analysis of Change in Implicit Bias*, RESEARCHGATE, May 2016, at 1, 32, https://www.researchgate.net/publication/308926636_A_Meta-Analysis_of_Change_in_Implicit_Bias.

39. *Id.* at 32.

40. Tom Bartlett, *Can We Really Measure Implicit Bias? Maybe Not*, CHRON. HIGHER EDUC. (Jan. 5, 2017), http://www.chronicle.com/article/Can-We-Really-Measure-Implicit/238807.

41. *See, e.g.*, Justin D. Levinson et al., *Guilty by Implicit Racial Bias: The Guilty/Not Guilty Implicit Association Test*, 8 OHIO ST. J. CRIM. L. 187, 196 (2010) ("Legal researchers often rely on the IAT for the proposition that people are implicitly biased, and tend to link it to a variety of legal claims.").

42. *Cf.* Jack Glaser et al., *Racial Bias and Public Policy*, 1 POL'Y INSIGHTS FROM BEHAV. & BRAIN SCI. 88, 90 (2014) ("The likelihood that much discrimination results from implicit biases and is therefore not intentional at the individual level has led legal scholars to call for changing how discrimination cases are litigated and adjudicated . . . .").

43. Proponents of the implicit-bias theory have defended their work against criticisms of this sort. *See, e.g.*, Anthony G. Greenwald et al., *Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects*, 108 J. PERSONALITY & SOC. PSYCHOL. 553, 557–58 (2015) ("Small effect sizes comprise significant discrimination. . . . Small effects can produce substantial discriminatory impact also by cumulating over repeated occurrences to the same person.").

take the problem of racial discrimination seriously, their strategizing toward reform should be firmly rooted in empirical reality. Even the most guarded interpretation of these new findings suggests a need for further investigation prior to continued action on this front.

Given these intersecting dynamics, the unfolding reproducibility crisis has called into question much of the behavioral science research upon which both policymaking and jurisprudence have been predicated. This problem isn't unique—the relationship of science and law has always been fraught with questions—but given the *scale* of the problem, and the existence of new, high-level commitments entrenching behavioral science in the world of law and policy, the need to address this crisis will only grow over time.

III. THE DEFERENCE DILEMMA

Much of this Essay's analysis centers on how federal agencies ought to more effectively use findings from behavioral science. This is particularly important in light of both the recent Executive Order entrenching it in agency decision-making and the unfolding scientific reproducibility crisis. But controversies over the role that such findings ought to play in governmental decision-making are quite longstanding: rather than simply being limited to internal agency deliber-

---

There are two problems with this defense, however. First, the Greenwald et al. paper (2015) predates the new meta-analysis conducted by Forscher et al. (2016); a greater amount of evidence is now available for review and analysis. *See id.*; Forscher et al., *supra* note 38. Second, the effect size anticipated—in Greenwald et al.'s own words, "more than 4% of variance in discrimination-relevant criterion measures is predicted by Black-White race IAT measures," Greenwald et al., *supra*, at 560—does not generally align with the sweeping rhetoric about implicit bias that other writers have employed. *See, e.g.*, German Lopez, *Why Police So Often See Unarmed Black Men as Threats*, VOX, http://www.vox.com/2014/8/28/6051971/police-implicit-bias-michael-brown-ferguson-missouri (last updated Sept. 20, 2016, 10:00 AM) (repeatedly describing the effects of implicit bias as "devastating.").

Contra Lopez and others, a great deal of scholarship has described racialized aspects of the criminal justice system in ways that reflect *explicit*, not *implicit*, racial bias. *See generally* NICOLE GONZALEZ VAN CLEVE, CROOK COUNTY: RACISM AND INJUSTICE IN AMERICA'S LARGEST CRIMINAL COURT (2016) (identifying manifestations of this explicit bias). Findings like these suggest that resources devoted to combating implicit bias could be directed elsewhere—perhaps toward more targeted intervention efforts—to more effectively fight discrimination.

ations, however, these earlier debates actually played out on far more visible stages.

A. The Rise of Daubert

Experts have long clashed over the proper role and use of behavioral science in courtrooms.[44] The net result of these controversies has been a sort of institutional détente; the existing *Daubert*-based legal regime governing the admission or exclusion of such expert opinion[45]—in the *courtroom* setting—is overwhelmingly deferential to purportedly scientific findings, including those arising from the behavioral science disciplines.[46] In some sense, this degree of institutional deference is entirely understandable from a theoretical, *ex ante* standpoint: neither judges nor juries are trained scientists with the capacity to independently analyze complex data and draw sound conclusions.[47] The structural constraints on the judicial process iden-

---

44. *See, e.g.*, Henry F. Fradella et al., *The Impact of* Daubert *on the Admissibility of Behavioral Science Testimony*, 30 PEPP. L. REV. 403, 405 (2003) [hereinafter Fradella et al., *Behavioral Science Testimony*] ("Since the time *Daubert* was decided, both courts and legal commentators have voiced concerns that *Daubert*'s focus on empirical testability, scientific falsifiability, and reliability and validity (including an assessment of error rates) may pose serious problems for expert testimony in the behavioral sciences."); Henry F. Fradella et al., *The Impact of* Daubert *on Forensic Science*, 31 PEPP. L. REV. 323, 334 (2004) ("[I]t might be a debatable point whether the forensic behavioral sciences constitute 'science' with[in] [Karl] Popper's definition of the term . . . .").

45. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593–94 (1993) ("Ordinarily, a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested. . . . Another pertinent consideration is whether the theory or technique has been subjected to peer review and publication. . . . Additionally, in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error . . . . Finally, 'general acceptance' can yet have a bearing on the inquiry.").

46. *See, e.g.*, Dale A. Nance, *Reliability and the Admissibility of Experts*, 34 SETON HALL L. REV. 191, 202 (2003) (describing *Daubert* as entrenching "deference to the norms of science, norms that must be applied directly by trial and appellate judges").

47. *See* David S. Caudill & Richard E. Redding, *Junk Philosophy of Science?: The Paradox of Expertise and Interdisciplinarity in Federal Courts*, 57 WASH. & LEE L. REV. 685, 764–65 (2000) ("[I]n cases involving scientific issues, courts cannot wait for scientists to resolve their own controversies. . . . [T]he Court

tified in *Daubert* have not changed; typically, courts do not independently order supplementary scientific research as a means of resolving the controversies set before them.[48] This reality is impossible to avoid.[49]

The Supreme Court has explained at length its rationale for a highly deferential regime concerning the admissibility of scientific testimony. Prior to today's *Daubert*-centric regime, the evidentiary framework of *Frye v. United States* stipulated that "while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs."[50] In other words, courts afforded immense deference to the professional consensus of scientists, without independently second-guessing the reliability of the methodologies involved.

Daubert v. Merrell Dow Pharmaceuticals, Inc.[51] changed the status quo—it required judges to determine whether scientific methodology had been applied reliably to the facts of a given proceeding—but preserved a degree of *Frye*'s deferential approach to expert testimony.[52] In *Daubert*, the Supreme Court interpreted the Federal

---

had to pick a winner, or at least try to pick a winner. The definition of science announced in *Daubert* was ambiguous. Thus, federal courts enjoy a certain philosophical leeway alongside and paralleling their leeway to make reliability determinations."); *see also* Christopher B. Mueller, Daubert *Asks the Right Questions: Now Appellate Courts Should Help Find the Right Answers*, 33 SETON HALL L. REV. 987, 1022 (2003) ("Critics have argued that judges cannot act constructively in the way that *Daubert* envisions, but there are good reasons to think that indeed judges can rise to the task. . . . [C]ourts are in fact working hard in very challenging areas to achieve appropriate outcomes in appraising science.").

48. *See supra* note 47.

49. It bears mention that the idea that courts are incapable of conducting research independently is not an entirely uncontested proposition. *See, e.g.*, Edward K. Cheng, *Independent Judicial Research in the* Daubert *Age*, 56 DUKE L.J. 1263, 1315 (2007) ("[I]ndependent research carries great promise as a tool for helping judges decide *Daubert* questions and for improving scientific decisionmaking [sic] in the courts generally.").

50. Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

51. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).

52. Graham R. Jones, *President's Editorial—The Changing Practice of Forensic Science*, 47 J. FORENSIC SCI. 437, 437 (2002) (noting that the "*Daubert* standard [went] a step further than *Frye*").

Rules of Evidence to afford lower courts significant leeway in admitting alleged experts, holding that:

> "General acceptance" is not a necessary precondition to the admissibility of scientific evidence under the Federal Rules of Evidence, but the Rules of Evidence—especially Rule 702—do assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand. Pertinent evidence based on scientifically valid principles will satisfy those demands.[53]

Six years later, in *Kumho Tire Co. v. Carmichael*, the Court reaffirmed the strong degree of latitude possessed by judges, reasoning that "[t]he trial court must have the same kind of latitude in deciding *how* to test an expert's reliability, and to decide whether or when special briefing or other proceedings are needed to investigate reliability, as it enjoys when it decides *whether or not* that expert's relevant testimony is reliable."[54] The *Kumho Tire* Court pushed the boundaries of institutional competence beyond *Daubert*, observing that whether the principles embedded in *Daubert* "are, or are not, reasonable measures of reliability in a particular case is a matter that the law grants the trial judge broad latitude to determine."[55] This "broad latitude" risks becoming laxity.[56]

---

53. *Daubert*, 509 U.S. at 597; *see also* Victor E. Schwartz & Cary Silverman, *The Draining of* Daubert *and the Recidivism of Junk Science in Federal and State Courts*, 35 HOFSTRA L. REV. 217, 244 (2006) ("As gatekeepers, judges have an obligation to keep theories out of the courtroom unless and until the expert's hypothesis is tested. Reliable expert testimony should not require a leap of faith.").

54. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 152 (1999).

55. *Id.* at 153.

56. *See* Jules Epstein, *Preferring the "Wise Man" to Science: The Failure of Courts and Non-Litigation Mechanisms to Demand Validity in Forensic Matching Testimony*, 20 WIDENER L. REV. 81, 96 (2014) (noting a "historic record of liberal admissibility of prosecution-offered forensic evidence."). Epstein goes on to decry the persistent failure of existing scientific review bodies to confront latent "limitations in their respective fields." *Id.* at 111.

Insights derived from behavioral science, for many judges, will suffice to meet the threshold requirements for testimonial admissibility.[57] In the words of one scholar:

> The social sciences most often find their way into the courtroom as a tool to account for or predict human behavior. The evidence usually consists of general assertions about classes of persons, such as rape victims, and is offered "to provide a social and psychological context in which the trier can understand and evaluate claims about the ultimate fact."[58]

Given this long-standing practice, it makes sense that a theory like implicit bias—whether or not implicit bias can be consistently connected to discriminatory conduct—would find purchase in courtrooms.[59] Moreover, wholly apart from the question of implicit bias evidence as an *evidentiary* matter, at least one judge has explicitly incorporated "implicit bias" instructions into jury charges as a means of pushing back against perceived patterns of discrimination in the criminal justice system.[60] The theory of implicit bias as a driver of discriminatory behavior, then, is rapidly migrating from the realm of academic debate into the domain of accepted cultural wisdom—even as the underlying science itself is coming into question.[61] And once this public entrenchment of this concept occurs, existing legal and political structures—including administrative agencies, which may have

57. *See* Mark S. Brodin, *Behavioral Science Evidence in the Age of* Daubert*: Reflections of a Skeptic*, 73 U. CIN. L. REV. 867, 869–70 (2005) ("[R]esearchers tracking *Daubert* have concluded that it has not resulted in significant changes in the admissibility of behavioral and social science evidence.").

58. *Id.* at 868 (internal citation omitted).

59. *See* Jerry Kang et al., *Implicit Bias in the Courtroom*, 59 UCLA L. REV. 1124 (2012) (arguing for broad application of implicit bias science in the courtroom context). *But see id.* at 1168 ("We concede that our claims about implicit bias influencing jury decisionmaking [sic] in civil cases are somewhat speculative and not well quantified.").

60. Bennett, *supra* note 34, at 391 n.a1.

61. *See* Alina Marciniak, *What Does It Take to Get Implicit Bias Training to Work?*, HR DAILY ADVISOR (July 11, 2018), https://hrdailyadvisor.blr.com/2018/07/11/take-get-implicit-bias-training-work/ ("Implicit bias training has become increasingly widespread after several highly publicized incidents of racial profiling.").

made significant internal decisions on the assumption that the concept is sound[62]—are not optimally suited to pushing back against an upstream methodological disruption that threatens to undermine pre-

62. To wit, a number of federal agency decisions and budgetary allocations are predicated upon the potentially flawed assumption that implicit bias predicts discriminatory behavior. For instance, in 2016, the Obama Administration institutionalized formal implicit bias training for Department of Justice personnel. Julia Edwards, *Justice Dept. Mandates 'Implicit Bias' Training for Agents, Lawyers*, REUTERS (June 27, 2016, 11:07 AM), http://www.reuters.com/article/us-usa-justice-bias-exclusive-idUSKCN0ZD251. While such training may well have merit independent of its relationship to potential discrimination, the rationale for its current implementation appears closely tied to ongoing allegations of perceived systemic discrimination.

But if no clear relationship actually exists between performance on implicit bias assessments and instances of racial discrimination, the program's methods are not properly aligned with its goals. The resources invested in implicit bias training might, for instance, be more constructively directed toward strategies of "training in human relations, of orientation to the cultures and subcultures of the people with whom criminal justice agents interact daily, and of supervisory oversight designed to detect and correct [overt] bias in the attitudes, speech, and behavior of subordinate personnel." SENTENCING PROJECT, REDUCING RACIAL DISPARITY IN THE CRIMINAL JUSTICE SYSTEM: A MANUAL FOR PRACTITIONERS AND POLICYMAKERS 9 (2008), http://www.sentencingproject.org/wp-content/uploads/2016/01/Reducing-Racial-Disparity-in-the-Criminal-Justice-System-A-Manual-for-Practitioners-and-Policymakers.pdf.

Similarly, the National Science Foundation has stressed the need for implicit bias training (and allocated resources to this area). *See* NAT'L SCI. FOUND., NATIONAL SCIENCE FOUNDATION'S DIVERSITY AND INCLUSION STRATEGIC PLAN 2012–2016 (IN SUPPORT OF THE GOVERNMENT-WIDE EFFORT TO ENHANCE DIVERSITY AND INCLUSION IN THE FEDERAL WORKFORCE) 13 (2012), https://nsf.gov/od/odi/reports/StrategicPlan.pdf. Given that agency resources are not unlimited, if *meaningfully* promoting diversity and inclusion is truly a goal to which the agency is committed, it likely makes more sense to allocate more resources to leadership programs that are oriented toward the retention and development of members of disadvantaged groups. The NSF's report further notes that its diversity and inclusion efforts include an "Aspiring Leader Program and a Senior Leadership Development Program, both of which will focus on identifying and developing leadership competencies across the full range of employees at NSF." *Id.* In lieu of implicit bias trainings that may not bear a strong relationship to patterns of entrenched discrimination in the field, such development programs may prove more effective.

viously accepted findings.[63] *Daubert* and *Kumho Tire* place admissibility questions in the hands of individual judges, and judges are likely to construe testimonial admissibility quite broadly.[64]

B. The Problem of Deference

Characterized in its broadest sense, this problem is not new, and some judicial deference to scientific expertise is obviously necessary. Judges and lawyers *aren't* scientists and cannot evaluate competing findings with the proficiency of a qualified expert. Some of the current ways in which such deference manifests, however, are resoundingly ill suited to addressing a broad crisis in scientific research methodology.[65] "[T]he law is often written to encourage experts to transcend the bounds of their professional knowledge,"[66] which forc-

63. *Cf.* Emily Hammond Meazell, *Super Deference, The Science Obsession, and Judicial Review as Translation of Agency Science*, 109 MICH. L. REV. 733, 752 (2011) (outlining the hypothesis that administrative agencies, by design, frequently suffer from a problem of accumulating useless "data" to support agency decisions without an eye to the reliability of that data).

64. *See, e.g.*, Sandra Guerra Thompson, *Judicial Gatekeeping of Police-Generated Witness Testimony*, 102 J. CRIM. L. & CRIMINOLOGY 329, 333 (2012) ("[T]rial courts generally have either been unwilling or unable to perform competent reliability screening in criminal cases.").

This tendency is cast into sharp relief by the problem of wrongful convictions. One scholar has proposed that judges recognize "'suspect evidentiary categories'— a few types of evidence that are both recurring features of wrongful convictions and not otherwise susceptible to correction through traditional trial mechanisms," to include "eyewitness identifications, confessions, forensic science, and jailhouse informant or snitch testimony." Keith A. Findley, *Judicial Gatekeeping of Suspect Evidence: Due Process and Evidentiary Rules in the Age of Innocence*, 47 GA. L. REV. 723, 726–27 (2013). This Essay is certainly not the first to advocate a narrowing by judges of the permitted range of expert testimony.

65. In light of *Daubert*'s failure to meaningfully restrict the introduction of spurious science, one scholar has urged lawyers to "turn to scientists who developed the scientific method as it applies to the science of the courtroom—biology, chemistry and simple physics to enlighten themselves—and substitute these approaches for the flawed and faulty premises advocated by *Daubert* and *Frye*." Barbara Pfeffer Billauer, Daubert *Debunked: A History of Legal Retrogression and the Need to Reassess "Scientific Admissibility*,*"* 21 SUFFOLK J. TRIAL & APP. ADVOC. 1, 2 (2016).

66. Daniel W. Shuman & Bruce D. Sales, *The Impact of* Daubert *and Its Progeny on the Admissibility of Behavioral and Social Science Evidence*, 5 PSYCHOL. PUB. POL'Y, & L. 3, 11 (1999).

es judicial actors to make threshold decisions about the inadmissibility or admissibility of particular testimony.[67] Those decisions hinge on questions of methodological credibility that are not properly delegated to experts themselves: a standard-form *Daubert* inquiry invoking "sufficient facts or data," "reliable principles and methods" accepted in the field, and reliable application of those methods to the facts in a case operates as little more than an empty mantra.[68] Most problematically for those unconvinced of its gatekeeping merits, the broad deference regime espoused in *Daubert* has not been confined to the federal context. *Daubert*-type norms have had a trickle-down effect throughout governmental organs,[69] a trend that dramatically raises both the theoretical and practical stakes.

Put simply, the logic of *Daubert* militates against independent reconsideration of propositions deemed "established" by scientific consensus, even though the process of reconsideration and reevaluation constitutes the very fabric of the scientific method. Where the scientific consensus changes—a phenomenon occurring now with respect to certain behavioral science findings—governmental institutions' existing notions of deference obstruct the possibility of policy adjustment.[70]

---

67. *See id.* at 9. ("Obviously, it would help if there were a common language between the courts and behavioral and social scientists. A common language would assist judges as educated consumers to understand the complexity of the admissibility decisions that they are making and be less likely to jump to conclusions that are only superficially thought through and potentially inaccurate.").

68. *See* David Crump, *The Trouble with* Daubert-Kumho*: Reconsidering the Supreme Court's Philosophy of Science*, 68 MO. L. REV. 1, 14–15 (2003). "[A] premature pronouncement that was intended to be flexible has become an established set of criteria. It was foolhardy for the Court to ignore what was going to happen, which was that trial judges would consider the four *Daubert* factors to be legal principles established by the Supreme Court." *Id.* at 40.

69. *See* Fradella et al., *Behavioral Science Testimony*, *supra* note 44, at 404–05 ("The impact of *Daubert*, however, is not limited to federal courts, since many states have also adopted the *Daubert* test for the admissibility of expert testimony."); Shuman & Sales, *supra* note 66, at 13 ("One of the interesting sequella of *Daubert* is determining to what extent the courts and society can use its teachings to instruct us about the appropriate use of behavioral and social science experts outside of the courtroom.").

70. At the court level, this problem is likely further compounded by *Chevron* deference and related doctrines. *See generally* Chevron, U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837 (1984).

But why might institutional deference to irreproducible science be so problematic? For one thing, from a purely idealistic standpoint, the government ought not entrench binding standards based on a "conventional wisdom" that does not map onto scientific reality.[71] But the problem is more severe than that.

Significant real-world harms follow from leaving this problem unaddressed. First, erroneous ideas about causality allow problems like discrimination to flourish unabated. Insofar as broad commitments to equality require conscious efforts against enduring legacies of discrimination and violence, holders of power have a moral responsibility to seek to achieve such goals.[72] The incorporation of bad science into policy fails the simple efficacy test and leaves the problem in place. Second, the misallocation of resources toward unproductive goals means that scarce finances are not being employed more strategically.[73] If system-wide commitments to effective use of taxpayer resources were taken seriously, stakeholders should seek to promptly redirect wasteful spending to more productive ends. Finally, the persistence of error risks inducing fatigue; other, more successful ways of challenging discriminatory habits undoubtedly exist, but a misguided fixation on implicit bias as a driver of discrimination likely "crowd[s] out" these other channels.[74]

These issues are not limited to the context of implicit bias but reflect problems within the larger doctrinal regime connecting science and law. In light of these looming risks, the next Part considers

---

71. *Cf.* ROBERT J. SAMUELSON, UNTRUTH: WHY THE CONVENTIONAL WISDOM IS (ALMOST ALWAYS) WRONG (2001) (exploring this problem in greater depth).

72. *See, e.g.*, Herman Finer, *Administrative Responsibility in Democratic Government*, 1 PUB. ADMIN. REV. 335, 350 (1941) ("Moral responsibility is likely to operate in direct proportion to the strictness and efficiency of political responsibility, and to fall away into all sorts of perversions when the latter is weakly enforced.").

73. *See, e.g.*, CASS R. SUNSTEIN, CONSPIRACY THEORIES AND OTHER DANGEROUS IDEAS 48 (2014) ("[P]ublic misfearing helps to produce significant misallocations of public resources.").

74. *Cf.* Sharon B. Jacobs, *The Administrative State's Passive Virtues*, 66 ADMIN. L. REV. 565, 579, 588 (2014) ("[S]tep-by-step regulation will allow the agency to enter the fray but to do so cautiously. It is therefore a helpful strategy for testing the waters and for making progress on an issue while avoiding regulatory fatigue.").

several ways governmental actors—either working in concert or, if that fails, working independently—might work to resist the potential encroachment of "bad science" within legal and political institutions.

IV. SKETCHING A STRATEGIC MULTI-BRANCH RESPONSE

The existing institutional architecture at the nexus of behavioral science and law is poorly positioned to respond to systemic methodological problems.[75] The existence of long-standing norms of deference and agency independence[76] means that stakeholders across all three spheres of government will need to adopt a somewhat coordinated response. In the antifragile policymaking model I propose, the initial impetus for strategic response would come from the executive branch and would be followed by actions within the legislative and judicial branches.

No single work can provide a solution guaranteed to succeed across the board, and this Essay offers merely a starting point for further investigation of this developing issue. Thus, these recommendations are necessarily limited in scope; they may, however, serve as a helpful foundation from which policymakers can more effectively interrogate the science that informs their decisions, building toward a culture of antifragility.

A. Executive Branch and Agency-Based Responses

With the ambitious goals of President Obama's initial executive order firmly in view, the Trump Administration should amend that order to explicitly direct agencies to adopt new safeguards against the entrenchment of unreliable science. These procedures should include ways to retrospectively evaluate, over time, the reproducibility of behavioral science research upon which initial agency

75. *Cf.* Levinson et al., *supra* note 41, at 187–88 ("Legal analysts have implicitly assumed that existing social cognition measures, many of which are carefully developed and rigorously tested (but not developed with the law in mind), are the only options for theory development in the legal context.").

76. *See, e.g.*, Jerry L. Mashaw, *Norms, Practices, and the Paradox of Deference: A Preliminary Inquiry into Agency Statutory Interpretation*, 57 ADMIN. L. REV. 501 (2005) (outlining these practices).

policy decisions are based.[77] Agencies should be directed to include, in budget proposals, requests for appropriations specifically directed to ongoing, internal research reassessments. These reassessments would conduct retrospective reviews of existing regulatory structures in light of ongoing developments within the behavioral science literature.[78] Some critics of perceived lawmaking excesses have called for "sunset provisions" that require that any decisions with legal force be reauthorized periodically according to a given interval;[79] echoing this model—but in lieu of adoption of its draconian approach—agencies should consider implementing formal "reanalysis triggers" that require formal cost-benefit analyses in light of potential evolutions in the underlying scientific field.[80]

---

77. This represents an expansion of the logic of Executive Order 12866, which requires that "[e]ach agency shall assess both the costs and the benefits of the intended regulation." Exec. Order No. 12,866, 58 Fed. Reg. 51,735 (Sept. 30, 1993). While initial agency decisions *to act* or *not to act* already take a broad range of circumstances and factors into account, the evaluation process should continue indefinitely.

78. The need for a discipline-specific response is grounded in the structural propensity of certain disciplines to report false-positive results. *See, e.g.*, Andrew Ferguson, *Making It All Up*, WKLY. STANDARD: ACCESS (Oct. 19, 2015, 12:00 AM), http://www.weeklystandard.com/making-it-all/article/1042807 ("Surveys have shown that published studies in social psychology are five times more likely to show positive results . . . than studies in the real sciences. This raises two possibilities. Either behavioral psychologists are the smartest researchers, and certainly the luckiest, in the history of science—or something is very wrong."); *see also* Joëlle Anne Moreno, *Einstein on the Bench?: Exposing What Judges Do Not Know About Science and Using Child Abuse Cases to Improve How Courts Evaluate Scientific Evidence*, 64 OHIO ST. L.J. 531, 534 (2003) ("As a first step, we should avoid the temptation to treat all science as a single field, which strips away meaning and practical value.").

79. *See, e.g.*, Frank H. Easterbrook et al., *Showcase Panel IV: A Federal Sunset Law*, 16 TEX. REV. L. & POL. 339, 340 (2012) (probing this question).

80. Generalized variations of this call to action are prevalent across the literature. *See, e.g.*, Cary Coglianese, *Moving Forward with Regulatory Lookback*, 30 YALE J. REG. 57, 64 (2013) ("Far too many of the retrospective reviews that agencies have conducted to date have been impressionistic, rather than systematic or rigorously empirical."). Contra Coglianese, however, the reanalysis triggers envisioned here would be predicated on the recognition that at present, certain scientific realms are *more likely to be undergoing disciplinary flux* and respond to that potential for methodological disruption accordingly. Regulations based on EPA groundwater testing, for instance, may not require the same sort of retrospective review *in*

A model reanalysis trigger might look something like the following: *every five years, any federal regulation substantially predicated upon findings from behavioral science literature must undergo a retrospective cost-benefit analysis that evaluates both regulatory performance and subsequent advances in the scientific field*. Both the Office of Information and Regulatory Affairs ("OIRA")[81] and the agency itself would be involved in this process: analyses of regulatory performance would be presented to decision makers alongside evaluations of developments in the behavioral science literature. The evidentiary "critical mass" required to reevaluate a decision or shift a policy might vary on a case-by-case basis or according to the determinations of individual agencies. In the face of a serious reproducibility crisis, then, policies based on erroneous or unproven research can be adjusted without continuing to suffer from a misalignment of goals and tactics. Through this process, agency regulations become more and more effective over time, contributing to a broader institutional culture of antifragile policymaking.[82]

---

light of current literature that might be needed for policies based on behavioral science.

81. *See Office of Mgmt. & Budget, Information and Regulatory Affairs*, WHITE HOUSE, https://www.whitehouse.gov/omb/information-regulatory-affairs/ (last visited Jan. 4, 2019).

82. The literature on the need for a form of "regulatory lookback" is extensive. *See, e.g.*, Melany C. Birdsong, *Reforming Regulation: No Time Like the Present*, 32 HAMLINE J. PUB. L. & POL'Y 371, 379 (2011) ("The failure to consistently implement rigorous monitoring and evaluation of regulations for the last four decades has resulted in the often inconsistent, inefficient and costly regulatory system in place today."); Reeve T. Bull, *Building a Framework for Governance: Retrospective Review and Rulemaking Petitions*, 67 ADMIN. L. REV. 265, 306–07 (2015) ("Absent any galvanizing event calling attention to a high profile regulatory failure, the combination of regulatory inertia and the endowment effect, with citizens reluctant to upset the prevailing regime, will generally prevent any major reassessment of the existing regulatory framework."); Jerry Ellig & Jerry Brito, *Toward a More Perfect Union: Regulatory Analysis and Performance Management*, 8 FLA. ST. U. BUS. REV. 1, 30–31 (2009); Matthew Wansley, *Cost-Benefit Analysis As a Commitment Device*, 87 TEMP. L. REV. 447, 458 (2015) ("Regulation is especially susceptible to obsolescence because its commands are so specific and detailed.").

This Essay's approach, however, is comparatively novel: not only would the model contemplated here expand the range of reanalysis sources contemplated within an agency's retrospective review, the regulatory lookbacks proposed would

To illustrate how reanalysis triggers might be successfully deployed, consider the following hypothetical scenario:

A new retirement-contribution system for federal employees requires that, when employees activate the program, they make a series of decisions about how to set up their contribution scheme. This is done through an online interface, NudgeNow, that presents a series of ostensibly intuitive choices. These NudgeNow choices affect the amounts of money that will be withheld from employees' bi-weekly paychecks.

The executive branch's SBST has played a key role in designing this online system and has done so in view of the behavioral science principle known as *ego depletion*. According to the ego depletion theory, the ability to make meaningful, reasoned choices—choices that require a degree of self-restraint—declines over time as more and more choices are presented successively.[83] In other words, ego depletion theory suggests that willpower can be drained over time.[84] When the NudgeNow interface is evaluated in light of the "ego fatigue" risk, the SBST realizes that the number of decisions employees must make at one time is simply too high. As a result of the fatiguing effect NudgeNow induces, the SBST believes that employees further along in the NudgeNow decision-making sequence will be less likely to make "responsible" decisions about paycheck withholdings and the allocation of their retirement contributions.[85] Here, "responsible" is understood as "that which benefits both the

be deployed primarily in domains where they are most needed—specifically, those policies underpinned by findings from behavioral science.

83. *See generally* Roy F. Baumeister, *Ego Depletion and Self-Control Failure: An Energy Model of the Self's Executive Function*, 1 SELF & IDENTITY 129 (2002) (introducing this concept).

84. *See generally* ROY F. BAUMEISTER & JOHN TIERNEY, WILLPOWER: REDISCOVERING THE GREATEST HUMAN STRENGTH (2011); Roy F. Baumeister et al., *Ego Depletion: Is the Active Self a Limited Resource?*, 74 PERSONALITY PROCESSES & INDIVIDUAL DIFFERENCES 1252 (1998) (outlining this theory).

85. Commentators have explored at length the difficulties associated with setting up retirement plans. *See, e.g.*, David V. Johnson, *Twilight of the Nudges*, NEW REPUBLIC (Oct. 27, 2016), https://newrepublic.com/article/138175/twilight-nudges ("[E]nrollment in retirement plans . . . is typically a complex and painful process. Since people are busy, don't like to incur cognitive costs, and often go along with the flow of life, many are liable not to bother to enroll at all, to their own detriment.").

employees and the government"—both parties, whether they realize it or not, have the goal of "increasing participants' retirement contributions through this plan" (to allow interest to accrue over time).

Accordingly, the SBST designs the program with an artificial stopping point partway through the NudgeNow sequence. The program pauses prior to full setup of the retirement-contribution scheme and displays a message that the "system is processing," and will require their attention at a point shortly thereafter.[86] Two to three days later, employees subsequently receive either a follow-up email or a phone call directing them to finish setting up the NudgeNow system. This, the SBST reasons, offsets the effect of ego depletion and allows the employees to make more rational decisions about their retirement plans.

New research suggests, however, that the ego depletion effect may be a victim of the reproducibility crisis.[87] Accordingly, in light of the fact that "ego depletion" may be a hypothesis without merit, the SBST's use of follow-up emails and phone calls—and the costs associated with them—may well prove to have been wasted efforts. The resources used to fund a psychological technique not rooted in scientific reality could have been employed more productively elsewhere (perhaps by making the NudgeNow software interface even easier and simpler to use).

---

86. *See, e.g.*, Kaveh Waddell, *Why Some Apps Use Fake Progress Bars*, THE ATLANTIC: TECH. (Feb. 21, 2017), https://www.theatlantic.com/technology/archive/2017/02/why-some-apps-use-fake-progress-bars/517233/?utm_source=atlfb (considering "a loan-approval app that builds suspense before delivering results to avoid making customers suspicious, and a site for delivering personalized phone-plan recommendations that slowed down its response time in order to convince users they were actually getting custom results").

87. *See, e.g.*, Martin S. Hagger & Nikos L.D. Chatzisarantis, *A Multilab Preregistered Replication of the Ego-Depletion Effect*, 11 PERSP. ON PSYCHOL. SCI. 546, 558 (2016) ("Results from the current multilab registered replication of the ego-depletion effect provide evidence that, if there is any effect, it is close to zero."); John H. Lurquin et al., *No Evidence of the Ego-Depletion Effect Across Task Characteristics and Individual Differences: A Pre-Registered Study*, 11 PLOS ONE 1, 15 (2016), http://dx.doi.org/10.1371/journal.pone.0147770 ("[W]e found no evidence of ego-depletion: participants in the Depletion Condition did not perform differently from participants in the Control Condition on the outcome task, contrary to the ego-depletion hypothesis.").

A reanalysis trigger in the original agency regulation authorizing this program's creation could have offset this risk. Under a system implementing a two-year reanalysis trigger, within two years of the retirement-contribution program's development and the initial rollout of NudgeNow, the SBST would have been required by OIRA to not only assess the effectiveness of the program on its own merits ("are people contributing?") *but also assess it with a view to the current behavioral science literature*. Such a reassessment would have alerted agency personnel to the research finding that "ego depletion" may not be a reproducible effect and that resources spent to combat it could be allocated elsewhere. This strategic approach to processing new findings would have allowed the SBST and other agency decision makers to rapidly react to new scientific developments and make the NudgeNow scheme even more effective—antifragile policymaking at its best.

In addition to the above-mentioned paradigm, wherever possible, executive branch decision makers can—and should—resist any tendency towards the gradual *Daubert*ization of institutional decision-making.[88] Despite the persistence of this doctrine in the legal realm, *Daubert*'s logic leaves important methodological issues underdeveloped, and it should not operate as an institutional norm for decision-making purposes.[89] Deference by courts might be necessary; deference to the status quo, by administrative agencies tasked with conducting their own fact-finding, should not be normalized.

88. I refer here to the process of organizations—in this case agencies—becoming overly deferential to experts and scientific findings.

89. The controversies over "regulatory *Daubert*" are highly salient. *See* David E. Bernstein, *What to Do About Federal Agency Science: Some Doubts About Regulatory* Daubert, 22 GEO. MASON L. REV. 549, 552 (2015) ("[F]ederal agencies have rejected appeals to implement *Daubert*-like standards when reviewing scientific evidence. On the other hand, a few decisions, mostly from the Seventh Circuit, have invoked the *Daubert* reliability test as informing their review of agency determinations, even while acknowledging that *Daubert* itself is not binding.").

Bernstein goes on to explain that regulatory *Daubert* "assumes that the underlying problem that needs to be addressed with agency decision making is with the unreliable science utilized by the agency, rather than with the regulatory standards established by the agency. In fact, the latter is often the source of discontent . . . ." *Id.* at 558–59. This Essay shares Bernstein's concern—hence, its call for reanalysis triggers under conditions likely to draw heated debates about evidential credibility.

Lastly, agencies charged with disbursing scientific research grants should consider prioritizing funding of research replication projects.[90]

B. Legislative Branch Responses

Legislators in Congress should support the aforementioned efforts to institutionalize reanalysis triggers[91] and finance further research replication within agencies. This would be a fundamentally bipartisan project: notwithstanding individual lawmakers' sentiments about the usefulness of agency actions *per se*, members of both major parties would likely agree on the importance of ensuring that federal agencies' decisions are based on sound evidence.[92]

If political consensus permits it, legislators should also consider revisiting Federal Rule of Evidence 702—which governs expert witness testimony—in light of the increasing methodological risks traced here.[93] One potential modification to Rule 702 might direct

---

90. *See* Amy Nussbaum, *ASA Advice for Funding Agencies on Reproducible Research?*, AM. STAT. ASS'N COMMUNITY (Oct. 12, 2016, 3:07 PM), http://community.amstat.org/blogs/amy-nussbaum/2016/10/12/asa-advice-for-funding-agencies-on-reproducible-research ("The funding model for reproducible research has not been worked out yet. . . . [The National Science Foundation] should sponsor research that evaluates various approaches to determining whether a finding replicates and to assess which approach(es) under which circumstances are the most helpful for reaching valid conclusions about replicability.").

91. In Texas, one promising legislative step forward was 2013's Senate Bill 344, which "allow[ed] courts to grant convicted individuals habeas corpus relief based on faulty or discredited scientific evidence." Naina Soni, *New Science, Old Convictions—Texas Senate Bill 344: Identifying Further Necessary Reform in Forensic Science*, 2 DUKE J.L. & BIOSCIENCES 149, 150 (2015). This incorporated a kind of scientific reanalysis trigger into the criminal justice apparatus, allowing prisoners to bring habeas actions based on post-conviction scientific developments calling into question the methodological principles upon which their convictions were predicated. Similar practices should be incorporated into the context of administrative agency policymaking.

92. *Cf.* Paul S. Miller & Bert W. Rein, *"Gatekeeping" Agency Reliance on Scientific and Technical Materials After* Daubert*: Ensuring Relevance and Reliability in the Administrative Process*, 17 TOURO L. REV. 297, 297–98 (2000) ("[A]gencies premise their actions on scientific and technical information, relying both on agency expertise and expert submissions from interested private parties. Courts reviewing these agency actions often defer to agency expertise on scientific and technical issues . . . .").

93. FED. R. EVID. 702.

post-*Daubert* courts to specifically consider, when evaluating the admissibility of expert testimony, whether the evidentiary findings presented by a given *social scientist* reflect conclusions shared across the field's literature.[94] Though courts may be loath to independently acknowledge the methodological divide between empirical and social sciences,[95] Congress is ideally situated to adjust the analytical ground (that is, the Federal Rules of Evidence) upon which admissibility decisions are ultimately predicated.[96]

Finally, in any legislation authorizing particular agencies to take particular actions, Congress should consider including as a matter of course "reproducibility provisions" that require any social science-based agency decisions to be justified via multiple research studies within comparable parameters.[97] These provisions would ex-

94. This proposal finds support in the existing literature. *See* Teresa S. Renaker, *Evidentiary Legerdemain: Deciding When* Daubert *Should Apply to Social Science Evidence*, 84 CALIF. L. REV. 1657, 1692 (1996) ("In order to preserve the integrity of the *Daubert* test as a tool for ensuring that only scientific testimony that meets the validity standard is admitted, courts should first determine whether proposed testimony functions as scientific or specialized knowledge, and only then assess scientific validity or helpfulness."); Cassandra H. Welch, Note, *Flexible Standards, Deferential Review:* Daubert*'s Legacy of Confusion*, 29 HARV. J.L. & PUB. POL'Y 1085, 1105 (2006) ("[T]he Court needs to apply a more active standard of review to evaluate whether expert testimony was appropriately reviewed and admitted or excluded and provide much-needed guidance by offering a more conservative standard for admissibility.").

95. *See, e.g.*, Brodin, *supra* note 57, at 943 ("It is not only *Daubert* but fundamental principles of evidence doctrine that demand considerably more skepticism than has been shown toward this mode of proof."). *But see* D.H. Kaye, *The Dynamics of* Daubert*: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 VA. L. REV. 1933, 2014 (2001) ("Phrases like 'gatekeeping' and 'intellectual rigor' are well and good, but heightened scrutiny should be reserved for methodology. Imperfections in the execution of a particular study should not result in exclusion unless . . . the probative value . . . is substantially outweighed by the dangers of prejudice, confusion, and time-consumption.").

96. *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 587 (1993) ("We interpret the legislatively enacted Federal Rules of Evidence as we would any statute.").

97. Similar procedures—that is, requirements that multiple evidentiary sources be employed—are followed within certain agency domains. *See* Matthew W. Swinehart, Note, *Remedying* Daubert*'s Inadequacy in Evaluating the Admissibility of Scientific Models Used in Environmental-Tort Litigation*, 86 TEX. L. REV. 1281, 1317–18 (2008) ("[F]or environmental-tort litigation, the EPA ranks certain

tend beyond traditional cost-benefit analyses (and even the reanalysis triggers envisioned above): their goal would not be to ascertain the most effective means of reaching a given goal but rather to field-test the implicit agency assumption that the proposed means *is likely to be effective at all.*[98]

C. Judicial Branch Responses

The reforms proposed above are no doubt ambitious, with potentially fraught consequences for the relationship between behavioral science researchers and legal institutions. Given current conditions of political gridlock and institutional dysfunction, the prospects of top-down institutionalization of reanalysis triggers—or legislative adjustments to Rule 702—might well appear bleak. And contemporary trend lines are not encouraging.

1. Judicial Self-Investigation

Former Attorney General Jeff Sessions declined to reauthorize the National Commission on Forensic Science, "a panel of judges, defense attorneys, researchers and law enforcement officials that had been advising the attorney general on the use of scientific evidence in the criminal justice process."[99] In so doing, Attorney General Sessions prematurely ended the Commission's "review of closed cases for inaccurate or unsupported statements by forensic analysts, which

models on the basis of their statistical performance in comparison to empirical data.").

98. In the wake of the ongoing methodological disruptions outlined in this Essay, members of the scientific community have proposed their own sets of best practices for future researchers. These best practices include blind studies, "rigorous training in statistics and research methods," "distributed data collection," oversight by third parties, and peer review at both pre-publication and post-publication stages. Marcus R. Munafò et al., *A Manifesto for Reproducible Science*, 1 NATURE HUM. BEHAV. 1, 3 (2017). In shaping the parameters for agency action in response to reproducibility concerns, Congress can and should draw on these accumulated insights. Indeed, the regulatory lookback mechanism this Essay proposes would constitute a form of discipline-specific, post-publication peer review.

99. Sadie Gurman, *Sessions' Justice Dep't Will End Forensic Science Commission*, ASSOCIATED PRESS (Apr. 10, 2017), https://www.apnews.com/c076ef99c48948e3856902cfca9e7b14.

regularly occur in fields as diverse as firearm and handwriting identification, and hair, fiber, shoe, bite mark and tire tread matching, and even fingerprinting analysis."[100] Terminating this process constitutes a dangerous setback to the ongoing dialogue surrounding existing evidentiary practices, particularly as the reproducibility crisis continues to impact scientific findings; at this juncture, policymakers and judges require more information, not less.

Assuming the executive branch's persistent inaction, courts may need to pursue independent ways of addressing the reproducibility crisis. At least one court has already done so, albeit in a context other than behavioral science review. In 2009, the New Jersey Supreme Court appointed a Special Master to evaluate the scientific merits of the state's evidentiary practices.[101] The Special Master's findings were sharply critical of the status quo—particularly regarding the reliability of eyewitness identification procedures[102]—and it seems entirely plausible that parallel analyses elsewhere would reveal similar problems, given that the flawed practices involved are employed "by 48 states and the federal courts."[103] If the executive branch is unable or unwilling to address existing evidentiary crises, courts must have the wherewithal to pick up the torch themselves and contribute to the public conversation on this subject.

2. Toward a Field-Based *Daubert*?

Judges are obviously positioned to affect an even broader change in the status quo. In his partial concurrence to *Daubert*, Justice Rehnquist contemplated the emerging challenge of judicial evi-

---

100. Erin E. Murphy, Opinion, *Sessions Is Wrong to Take Science Out of Forensic Science*, N.Y. TIMES (Apr. 11, 2017), https://www.nytimes.com/2017/04/11/opinion/sessions-is-wrong-to-take-science-out-of-forensic-science.html.

101. *See* Alana Salzberg, *Special Master Appointed by N.J. Supreme Court Calls for Major Overhaul of Legal Standards for Eyewitness Testimony*, INNOCENCE PROJECT (June 21, 2010), https://www.innocenceproject.org/special-master-appointed-by-n-j-supreme-court-calls-for-major-overhaul-of-legal-standards-for-eyewitness-testimony/.

102. *Id.*

103. *Id.*; *see also* COMM. ON IDENTIFYING THE NEEDS OF THE FORENSIC SCIS. CMTY., NAT'L RESEARCH COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD (2009) (broadly identifying the scientific challenges facing currently accepted American forensic science).

dentiary gatekeeping, commenting that the case's "unusual subject matter [that is, scientific credibility] should cause us to proceed with great caution in deciding more than we have to, because our reach can so easily exceed our grasp."[104] Justice Rehnquist was accordingly reticent to share the Court's "confidence that federal judges can make a 'preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue.'"[105] As today's dilemmas illustrate, Rehnquist's concern was prescient. *Daubert*'s two-pronged test—encompassing both concern for scientific validity and proper application in context—implicitly incorporates *Frye*-type deference to an abstracted consensus of scientists.[106] The key problem with a highly deferential *Frye*-type approach, however, is that it offers no opportunity for judicial branch actors to help counter upstream methodological problems. Thus, *Daubert*'s existing deference regime places undue weight on a "scientific consensus" that may well be seriously flawed.

Jurisprudentially, one possible solution to this dilemma might be an organic doctrinal evolution away from a *case*-based *Daubert*—an evidentiary approach that focuses on the application of scientific methodology to a given set of facts and treats "scientific methodology" in a highly deferential way—toward a *field*-based *Daubert* that embraces field-specific standards for evaluating evidence. In other words, judges should focus less on individual situations and cases and more on the *types of evidence* being introduced. This would complete the shift away from *Frye*'s regime of maximal deference to "consensus" and allow judges to weigh the relative credibility of different evidentiary categories.

Obviously, any such shift toward a field-based *Daubert* would produce many inter-institutional tensions and tradeoffs. Some advantages of this approach might include: courts' increased agility in responding to systematic reliability failures within particular scientific disciplines, their capacity to meaningfully act to preserve litigants' rights even in the face of political gridlock, and the potentially

104. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 599 (1993).

105. *Id.* at 600.

106. *Id.* at 598.

catalyzing effect that concerted exercises of judicial discretion could have on policymakers within the political branches of government.

Potential challenges and disadvantages of a field-based *Daubert*, however, are also readily identified. A field-based *Daubert* would still probably be at least somewhat dependent on expert opinion—that is, expert opinion identifying the existence of a problem warranting additional judicial scrutiny. This dependency means that a narrow line exists between expert "meta-opinion" to which a judge *ought* defer—that is, the expert opinion calling into question the credibility of experts in other fields—and expert opinion within a particular field (say, forensic behavioral sciences) to which a judge *ought not* defer uncritically. This seems to introduce a new element of arbitrariness into the existing evidentiary regime.

Upon closer examination, however, this "objection from imprecision" is not insurmountable. In today's status quo, judges need not adopt a categorical, all-or-nothing approach to expert deference, and they need not start doing so here. Insofar as critical, discipline-specific meta-analyses play an important role in how science advances, judges may—within the latitude afforded them—take notice of whether a particular field of evidence is, on net, more prejudicial than probative.[107] They may thereby determine whether to admit evidence from such a field—all things considered—into a given case. Judges are vested with discretion, and expert deference need not be blind.

In short, under a field-based *Daubert*, judges presiding over trials would need to carefully weigh the prudence of adopting a case-by-case approach to evaluating evidence (or subcategories of evidence) rather than deferring to self-professed experts across-the-board. Ideally, a broader-scale collective approach to engaging these issues would be spearheaded by other branches of government—chiefly, the executive and its agencies, which have already demonstrated a willingness to differentiate between discrete scientific fields.[108] Failing that, however, judges themselves may be positioned

107. *Cf.* FED. R. EVID. 403. It warrants mention that some precedent for this kind of line drawing already exists. For instance, following the report of the Special Master, the New Jersey Supreme Court concluded that "in rare cases, judges may use their discretion to redact parts of identification testimony, consistent with [Federal Rule of Evidence] Rule 403." State v. Henderson, 27 A.3d 872, 925 (N.J. 2011), *modified,* State v. Chen, 27 A.3d 930, 942–43 (N.J. 2011).

108. *See, e.g.*, *Executive Order*, *supra* note 11.

to draw these important lines.[109] Consistent judicial enforcement of something like a field-based *Daubert*, even if not officially codified, could eventually develop into a system-wide norm.

Under ideal circumstances, the Supreme Court would revisit the methodological issues of *Daubert* and *Kumho Tire* and adopt a field-based standard more properly suited to contemporary evidentiary challenges. Such a standard might then, like the two-pronged *Daubert* norm of today, filter through federal and state organs and help build more robust system norms. As useful as such a reevaluation might be, however, strong institutional pressures—namely, the massive corpus of case law that has developed in *Daubert*'s wake[110]—militate against such reconsideration.

3. Incremental Reforms

Whether such higher-level action becomes reality, individual actors within the judicial system can immediately take certain concrete steps to mitigate the effects of the upstream scientific crisis.[111] Judges should resist the temptation to incorporate their preferred theories into juror instructions—or, if they do so, should be ready to adjust their practices based on subsequent findings.[112] Lawyers should be generally aware of the reproducibility crisis's potential impact on

---

109. *Cf.* Alex Kozinski, *Criminal Law 2.0*, 44 GEO. L.J. ANN. REV. CRIM. PROC. iii, xxxv (2015) ("[C]ourts must be far more rigorous in enforcing *Daubert* before allowing experts to testify in criminal trials.").

110. *See, e.g.*, Andrew W. Kane, *Basic Concepts in Psychology and Law*, *in* CAUSALITY OF PSYCHOLOGICAL INJURY: PRESENTING EVIDENCE IN COURT 261, 275 (Gerald Young et al. eds., 2007) ("[A]s of April 16, 2006, there had been 761 federal appellate court decisions on *Daubert* issues since January 1, 2000.").

111. For a more comprehensive set of proposed reforms to judicial practice, see Maxine D. Goodman, *A Hedgehog on the Witness Stand—What's the Big Idea?: The Challenges of Using* Daubert *to Assess Social Science and Nonscientific Testimony*, 59 AM. U. L. REV. 635, 677–82 (2010).

112. As it were, the converse of this problem—that is, calling for juries to display *greater skepticism* of certain evidence, rather than place *greater weight* on certain social-scientific theories—could play a helpful role in pushing back against scientific unreliability. Indeed, the New Jersey Supreme Court, in keeping with its aforementioned interest in evidentiary integrity, has crafted its own judicial counterweight to the risk of unreliable evidence, requiring that "enhanced instructions be given to guide juries about the various factors that may affect the reliability of an identification in a particular case." *Henderson*, 27 A.3d at 924.

evidentiary credibility and rigorously cross-examine experts whose work might prove potentially subject to the truth-blurring effects of non-reproducibility. Jurors should familiarize themselves with the underlying methodological problem and treat attorneys' invocation of behavioral science-based conclusions with suitable caution.[113] And wherever opportunities present themselves, state-level courts should follow New Jersey's successful model and develop independent sets of best practices for grappling with evidentiary science, which can then be implemented across lower levels of the state judiciary.

V. CONCLUSION

The developing integration of behavioral science with law and policy has engendered both promise and peril. Early hopes for more effective government action must be weighed against the creeping risk of defective methodology—a risk that ultimately bears on the administration of justice, and a risk that traditional principles of institutional deference are poorly adapted to counter. The existing *Daubert* framework rests on a creaky foundation.

A systematic and coordinated institutional answer, involving all three branches of government and using reanalysis triggers where possible, would likely prove the most powerful antidote to the triple threat of unjust procedural outcomes, system-wide delegitimization, and epistemic nihilism. Failing that, individual branches of the government—and individual members of those branches, particularly judges with the will to act—can still take steps to resist the structural reproducibility problems afflicting modern social science and promote antifragility. Where possible, they should do so. In the words of Cass Sunstein himself, "[w]hat is needed is a genuine culture of retrospective analysis, in which agencies stand ready and willing to improve and simplify rules completed decades ago, or years ago, or months ago, or even weeks ago. Well-functioning companies are

113. There are grounds to believe that jurors are already likely to hold evidence drawn from behavioral science to a higher standard. *See* Richard J. Bonnie & Christopher Slobogin, *The Role of Mental Health Professionals in the Criminal Process: The Case for Informed Speculation*, 66 VA. L. REV. 427, 465 n.121 (1980) ("We think laymen are naturally skeptical about the scientific nature of psychiatric and psychological expertise . . . . [W]e feel the risk of 'expert dominance' is grossly exaggerated.").

flexible and adaptive. . . . The same should be true of government."[114]

Ultimately, if behavioral science-based policymaking is to remain viable for the long term, it must stanch the gaping conceptual wound that has opened up at its base. Deploying stronger institutional checks on irreproducible scientific findings—and in so doing, advancing a future-minded norm of antifragile policymaking—would be an excellent start.

114. Cass R. Sunstein, *The Regulatory Lookback*, 94 B.U. L. REV. 579, 602 (2014).