ISLANDS OF ALGORITHMIC INTEGRITY: IMAGINING A DEMOCRATIC DIGITAL PUBLIC SPHERE

AZIZ Z. HUQ*

TABLE OF CONTENTS

* Frank and Bernice J. Greenberg Professor of Law, University of Chicago Law School, and associate professor, Department of Sociology. Thanks to Erin Miller for extensive and illuminating comments, and to participants in the symposium—in particular Yasmin Dawood—for terrific questions and conversation. The editors of the *Southern California Law Review*, in particular Michelle Solarczyk and Tyler Young, did exemplary work in making this essay better. The Frank J. Cicero Foundation provided support for this work.

INTRODUCTION

A class of digitally mediated online platforms play a growing role as the primary sources of Americans' knowledge about current events and politics. Prominent examples include Facebook, Instagram, TikTok, and X (which had formerly been known as Twitter). While only eighteen percent of Americans cited social media platforms as their preferred source of news in 2024, this number had risen by a striking six points since 2023.[1] These platforms also compete in "one of the most concentrated markets in the United States,"[2] as a consequence of network effects and high barriers to entry.[3] Current trends suggest that social media will soon outpace traditional news websites as the main source for a plurality of Americans' understanding of what happens in the world.[4] Such platforms, which I will call "social platforms" here, are thus in practice a central plank of the political public sphere given their growing role in supplying so many people with news.

The role that social platforms play in public life has sparked a small avalanche of worries even before the extraordinary entanglement of big tech's corporate leadership with the partisan policy projects of the second Trump administration.[5] The worries are diverse. Many commentators have aired concerns about the effects of social-platform use on mental health and sexual mores,[6] or the extent of economic exploitation in this platform-based gig economy.[7] These important cultural and economic worries are somewhat

1. Christopher St. Aubin & Jacob Liedke, *News Platform Fact Sheet*, PEW RSCH. CTR. (Sept. 17, 2024), https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet [https://perma.cc/SJ49-28W6].

2. Caitlin Chin-Rothmann, *Meta's Threads: Effects on Competition in Social Media Markets*, CTR. FOR STRATEGIC & INT'L STUD. (July 19, 2023), https://www.csis.org/analysis/metas-threads-effects-competition-social-media-markets [https://perma.cc/2MQN-YSUR].

3. *Id.*

4. St. Aubin & Liedke, *supra* note 1.

5. This essay was completed in late 2024 and edited in early 2025. I have not tried here to account for the synergistic entanglement of Elon Musk and the Trump White House, nor for the ways in which the X social platform has changed as a result. It is, as I write, too early to say how this exorbitant display of codependency between partisan and technological projects will alter the American public sphere.

6. *See, e.g.*, *Surgeon General Issues New Advisory About Effects Social Media Use Has on Youth Mental Health*, U.S. DEPT. OF HEALTH & HUMAN SERVS. (May 23, 2023), https://www.hhs.gov/about/news/2023/05/23/surgeon-general-issues-new-advisory-about-effects-social-media-use-has-youth-mental-health.html (noting "ample indicators that social media can also pose a risk of harm to the mental health and well-being of children and adolescents").

7. *See, e.g.*, Veena Dubal, *On Algorithmic Wage Discrimination*, 123 COLUM. L. REV. 1929, 1944 (2023).

distinct from worries surrounding the political functions of the digital public sphere. It is the latter's pathologies, and only those problems, that this essay—as well as the broader symposium on listeners' rights in which it participates—concentrates on.

Even within the narrower compass of political speech defined in strict and demotic terms, the role of social platforms raises several distinct concerns. I take up three common lines of criticism and concern here. A first line of critique focuses on these platforms' alleged harmful effects on a broad set of user beliefs and dispositions thought to be needful for democratic life. Social platforms, it is said, pull apart the electorate by feeding them fake news, fostering filter bubbles, and foreclosing dialogue—to the point where democratic dysfunction drives the nation toward a violent precipice. This first argument concerns platforms' effects on the public at large.

A second common line of argument, by contrast, makes no claim about the median social platform user. It instead focuses on the "radicaliz[ing]" effect of social media engagement on a small handful of users at the ideological margin.[8] If even these few users resort to violence to advance their views, it might be said that social media has had a deadly effect.[9] This is an argument not about social platforms' effects on the mass of users, but upon the behavior of a small tail of participants in the online world.

Yet a third sort of argument against social platforms does not sound in a strictly consequentialist register. It does not lean, that is, on any empirical evidence as to how users are changed by their engagement. Rather, it is a moral argument that picks out objectionable features of the relationship between platforms and their users. This plainly asymmetrical arrangement, it is said, allows invidious manipulation, exploitation, or even a species of domination. Even if users' behaviors do not change, these characteristics of the platform-user relationship are said to be insalubrious. Especially given the role that algorithmic design plays in shaping users' online experiences, it is argued, a morally problematic imbalance emerges between ordinary people and the companies that manage social platforms. In the limited case, in which there are few potential sources of information and in which those sources are controlled and even manipulated by their owners (usually men of a certain age who are disdainful of civility and truthfulness norms), an acute concern about domination arises.

8. Steven Lee Myers & Stuart A. Thompson, *Racist and Violent Ideas Jump from Web's Fringes to Mainstream Sites*, N.Y. TIMES (June 1, 2022), https://www.nytimes.com/2022/06/01/technology/fringe-mainstream-social-media.html [https://web.archive.org/web/20250219041047/https://www.nytimes.com/2022/06/01/technology/fringe-mainstream-social-media.html].

9. *Id.*

If one accepts one of these arguments (and I will try to offer both their best versions and to explore their weaknesses in what follows), then there is some reason to think closely about the way social platforms are governed, and to look for regulatory interventions. Such governance might be supplied by platforms' own endogenous rules, which are usually embodied in their contractual terms of service or other internal procedures (such as mechanisms to dispute a take-down or deplatforming decision). Alternatively, governance could be supplied by exogenous legislation or regulation promulgated by a state. Private governance and legal regulation, of course, are potential substitutes. They can both be used to achieve the same policy goals. But how? What should such governance efforts, whether private or public, aspire to? And which policy levers are available to achieve it?

Where a platform employs algorithmic tools to shape users' experience by determining what they see, the range of potential interventions will be especially large. This is a result of the complexity of common computational architectures today. There are many ways to craft the algorithms on which many platforms run.[10] And there are many technical choices about which instruments to use, how to calibrate them, and what parameter (engagement? a subset of engagement?) to optimize. Many of these decision points offer opportunities for unavoidably normative choices about the purpose and intended effects of social platforms. Resolving those choices in turn requires some account of what it means exactly to talk about a normatively desirable social platform: That is, what should a social platform do? And for whom?

Such questions takes on greater weight given (1) recent regulatory moves by American states to control platforms' content moderation decisions;[11] (2) a recent Supreme Court decision responding to those efforts;[12] and (3) the European Union's Digital Services Act, a statute that takes yet a different and more indirect tack in modulating platform design and its ensuing costs.[13] Or consider a 2025 U.S. Supreme Court decision, rendered on a tightly expedited schedule, to uphold federal legislation

10. *See* Arvind Narayanan, *Understanding Social Media Recommendation Algorithms*, KNIGHT FIRST AMEND. INST. 9–12 (March 9, 2023), https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms [https://perma.cc/9WVD-7NJ6] (discussing common structural elements).

11. Tyler Breland Valeska, *Speech Balkanization*, 65 B.C. L. REV. 903, 905 (2024) ("In 2021 and 2022 alone, state legislators from thirty-four states introduced more than one hundred laws seeking to regulate how platforms moderate user content.").

12. Moody v. NetChoice, LLC, 603 U.S. 707 (2024); *see infra* text accompanying notes 124–26.

13. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 3 [hereinafter "Digital Services Act"].

banning TikTok.[14] The decision makes the remarkable suggestion that legislative control over social platforms—exercised by reshaping (or cutting off) the ordinary market from corporate control (for example, by forcing or by restricting a sale)—raises only weak First Amendment concerns. Applied broadly, such an exception from close constitutional scrutiny might allow broad state control over social platforms.

My main aim in this essay is to offer a new and fruitful analytic lens for thinking about these problems as questions of democratic institutional design. This is a way of approaching the problem of institutional design, not a set of prescriptions for how to do such design. I do so by pointing to a model of a desirable platform, and then asking how we can move toward that aspiration, and how much movement might be impeded or even thwarted. My aspirational model is not conjured out of the ether; rather, I take inspiration from an idea found in the scholarly literatures in political science and sociology that evaluates pathways of economic development. The idea upon which I draw is that development policy should aim to seed "islands of integrity" into patrimonial or nepotistic state structures as a way of building foundations for a more robust—and hence public-regarding—state apparatus.[15] This literature focuses on the question of the state's seeds and nurtures zones (or those of another interested party, such as a private foundation or an international organization) where public-regarding norms, not self-regarding or selfish motives, dominate as a means of generating public goods.

By analogy to the examples of effective public administration discussed in this literature, I will suggest here that we should think about public-regarding platforms as "islands of algorithmic integrity" that advance epistemic and deliberative public goods with due regard to the potential for either exploitation or manipulation inherent in the use of sophisticated computational tools. With that threshold understanding in mind, we should then focus on how to achieve that specific, affirmative model—and not simply on how to avoid narrowly-defined and specific platform-related

14. TikTok Inc. v. Garland, 145 S. Ct. 57, 72 (2025) (per curiam). The legislation in question is the Protecting Americans from Foreign Adversary Controlled Applications Act, Pub. L. No. 118–50, 138 Stat. 955 (2024).

15. For examples of the term in recent studies, see Monica Prasad, *Proto-Bureaucracies*, 9 SOCIO. SCI. 374, 376 (2022); Eliška Drápalová & Fabrizio Di Mascio, *Islands of Good Government: Explaining Successful Corruption Control in Two Spanish Cities*, 8 POL. & GOVERNANCE 128, 128 (2020). For further discussion, see *infra* Part II.

harms. An affirmative ideal, that is, provides a baseline against which potential reform proposals can be evaluated.[16]

To be very clear up front, this approach has limitations. It draws on the "island of integrity" literature here as a general source for inspiration, instead of a source for models that can be directly transposed. I do not think that there is any mechanical way of taking the lessons of development studies and applying them to the quite different virtual environment of social platforms. To the extent lessons emerge, they are at a high level of abstraction. Still, studies of islands of bureaucratic integrity in the wild can nevertheless offer a useful set of analogies: they point toward the possibility of parallel formations in the online world. They also help us see that there are already significant web-based entities that exemplify certain ideals of algorithmic integrity in practice because they hew to the general lessons falling out of the islands of integrity literature. These studies can illuminate how a more democratically fruitful digital public sphere might begin to be built given our present situation, even if they cannot offer a full blueprint of its ultimate design.

It is worth noting that my analytic approach here rests on an important and controversial assumption. That is, I help myself to the premise that reform of the digital public sphere can proceed first by the cultivation of small-scale sites of healthy democratic engagement and that these can be scaled up. But this assumption may not be feasible. It may instead be necessary to start with a "big bang": a dramatic and comprehensive sweep of extant arrangements followed by a completely new architecture of digital space. If, for example, you thought that the problem of social platforms began and ended in their concentrated ownership in the hands of a few bad-spirited people, then the creation of new, more democratic platforms would not necessarily lead to a comprehensive solution. Given disagreement about the basic diagnosis of social platforms' malady, it is hard to know which of these approaches is more sensible. Therefore, there is some value to exploring a piecemeal reform approach of the sort illuminated here. But that does not rule out the thought that a more robust "big bang" approach is in truth needed.

Part I of this essay begins with a brief survey of the main normative (consequentialist and deontic) critiques that are commonly lodged against social platforms, focusing on the three listed above. In Part II, I introduce the "islands of integrity" lens—briefly summarizing relevant sociological and

16. I am hence not concerned here with the First Amendment as a template or limit to institutional design. The constitutional jurisprudence of free speech provides a different benchmark for reform. I largely bracket that body of precedent here in favor of an analytic focus on the question of what functionally might be most desirable.

political science literature—as a means to directly think about social platform reforms. My aim in so doing is to provide a litmus test for thinking about social platform reform in the round. With that lens in hand, Part III critically considers the regulatory strategies pursued by the American states and the European Union to date. I suggest some reasons to worry that these are unlikely to advance islands of algorithmic integrity. I close by reflecting on some alternative regulatory tactics that might move us quicker toward that goal.

I. THE CASE(S) AGAINST SOCIAL PLATFORMS

What is a social platform? Do such all platforms work in the same way and raise the same kind of normative objections? Or are objections to platforms better understood as training on a subset of cases or applications? This Part sets some groundwork for answering these questions by defining the object of my inquiries and by offering some technical details about different kinds of platforms. I then taxonomize the three different objections that are commonly lodged against social platforms as they currently operate.

A. DEFINING SOCIAL PLATFORMS AND THEIR ALGORITHMS

A "platform" is "a discrete and dynamic arrangement defined by a particular combination of socio-technical and capitalist business practices."[17] A subset of platforms are understood by their users as distinctively "social" rather than "commercial" insofar they provide a space for interpersonal interaction, intercalated with other activities such as "reading political news, watching media events, and browsing fashion lines."[18] The leading "social platforms," as I shall call them here, are Facebook, X, Instagram, and TikTok.[19]

Not all social platforms propagate content in the same way. There are two dominant kinds of system architecture. The first is the social network, where users see posts by other users who they follow (or subscribe to) as well as posts those users chose to amplify.[20] When Facebook and Twitter allowed users to reshare or retweet posts, they enabled the emergence of networks of this sort.[21] Here, what one sees depends on who one "knows."

17. Paul Langley & Andrew Leyshon, *Platform Capitalism: The Intermediation and Capitalisation of Digital Economic Circulation*, 3 FIN. & SOC'Y 11, 13 (2017).

18. Lisa Rhee, Joseph B. Bayer, David S. Lee & Ozan Kuru, *Social by Definition: How Users Define Social Platforms and Why It Matters*, TELEMATICS & INFORMATICS, 1, 1 (2020).

19. *Id.* I have added TikTok to the list in the cited text. I use the term "social platforms" because "social media platforms" is overly clunky and merely "platforms" is too vague.

20. Narayanan, *supra* note 10, at 10.

21. *Id.* Note that before the affordances that allowed users to share content in these ways, these had limited network capacity.

Interconnected webs of users on a network can experience "information cascades" as information flows rapidly across the system.[22] This is known colloquially as "going viral." The possibility of virality depends not just on platform design but also on users' behaviors. But, in practice a very small number of posts go viral on social networks.[23] Attention is a scarce commodity. We cannot and do not absorb most of what's posted online. Our inability to absorb much means that it is only possible for a few items to achieve virality.

The second possible architecture is centered around an algorithm (or, more accurately, algorithms). On platforms of this sort, the stream of data observed by a user is largely shaped by a suite of complex algorithms, which are computational decisional tools that proceed through a series of steps to solve a problem. These algorithms, in the aggregate, are designed with certain goals in mind, such as maximizing the time users spend on the platform.[24] TikTok's "For You Page," Google Discover, and YouTube all rely at least in part on algorithms.[25]

In practice, what is for the sake of simplicity called "the algorithm" can be disaggregated into several different design elements, each of which is in truth a distinct algorithm or digital artifact. These include (1) the "surfaces of exposure" (that is, the visual interface encountered by users); (2) a primary ranking model (often a two-stage recommender system that combs through and filters potential posts); (3) peripheral models, which rank content that appears around the main surface of exposure (for example, ads); and (4) auxiliary models (for example, content moderation for illegal materials or posts that violate terms of service).[26] For the sake of simplicity, I will refer to them together only as "the algorithm," but it is worth keeping in mind that this is a simplification, and in fact there are multiple instruments at stake.

Algorithm design implicates many choices. At the top level, for example, an algorithmic model can be braided into a network model or integrated into a subscription-service model.[27] At a more granular level,

22. *Id.*

23. *Id.* at 15.

24. *Id.* at 10. Networks require both content processing tools (e.g., face recognition, transcription, and image filters) and also content propagation tools (e.g., search, recommendation, and content moderation). *Id.* at 8. I am largely concerned here with content propagation tools.

25. *Id.* at 11.

26. Kristian Lum & Tomo Lazovich, *The Myth of the Algorithm: A System-Level View of Algorithmic Amplification*, KNIGHT FIRST AMEND. INST. (Sept. 13, 2023), https://knightcolumbia.org/content/the-myth-of-the-algorithm-a-system-level-view-of-algorithmic-amplification [https://perma.cc/4WBQ-34WN].

27. Narayanan, *supra* note 10, at 10–11 ("[N]o platform implements a purely algorithmic model . . . .").

algorithms can be designed to optimize a broad range of varied parameters. These range from "meaningful social interactions" (Facebook's measure at one point in time) to user's watch time (YouTube's measure) to a combination of liking, commenting, and watching frequencies (TikTok's measure).[28] The choice of parameter to optimize is important. Most common parameters quantify some element of users' engagement with the platform, but they do so in different ways. Engagement measures are relevant from the platforms' perspectives given their economic reliance on the revenue from advertising displayed to users.[29] In theory, more engagement means more advertising revenue. But engagement on social platforms is surprisingly sparse. Somewhere between only one percent and five percent of posts on most social platforms generate any engagement at all.[30] And the movement from engagement to advertising is rarer still: most targeted online advertising is simply "ignored."[31]

B. CONSEQUENTIALIST CRITIQUES OF SOCIAL PLATFORMS

There are, as I read the literature, three clusters of normative concerns raised by social platforms that merit consideration as the most important and common criticisms made of those technologies.[32] Two are consequentialist, in the sense of training on allegedly undesirable effects of social platforms. Of course, such arguments need some means of evaluating downstream effects as either desirable or undesirable. In practice, they rest on some account of democracy as an attractive—even ideal—political order. (Note that as is often the case in legal scholarship, the precise kind of "democracy" at work in these critiques is not always fully specified. This lack of specification is a gap that will prove relevant in the analysis that follows.)[33] The other cluster is deontic, in the sense of picking out *intrinsically* unattractive qualities of social platforms. These accounts do not rely on a causal claim about the effects of social platforms; they instead assert the prima facie unacceptability of platforms in themselves.

Let's begin with the two consequentialist arguments and then move on to the deontic critique.

28. *Id.* at 19.

29. For a useful account of the behavioral advertising industry, see generally TIM HWANG, SUBPRIME ATTENTION CRISIS (2020).

30. Narayanan, *supra* note 10, at 28.

31. HWANG, *supra* note 29, at 77; *accord* Narayanan, *supra* note 10, at 29.

32. I recognize that there are complaints beyond those that I adumbrate here. I have selected those that seem to me supported by evidence and a coherent moral theory. I have ignored those wanting in such necessary ballast.

33. For an illuminating recent discussion on the varieties of democratic theory, see generally JASON BRENNAN & HÉLÈNE LANDEMORE, DEBATING DEMOCRACY: DO WE NEED MORE OR LESS? (2021).

A first view widely held in both the academic and non-academic public spheres is that social platforms cause political dysfunction in a democracy because of their effects on the dispositions and beliefs of the general public.[34] Using social platforms, this argument goes, drives (1) a dynamic of "affective polarization" (negative emotional attitudes towards members of opposition parties), or (2) traps us in "echo chambers" or filter bubbles that are characterized by limited, biased information.[35] Social media users are also said to be exposed to "fake news," which are "fabricated information that mimics news media content in form but not in organizational process or intent."[36] High levels of exposure are said to be driven by algorithmic amplification.[37] Recent advances in deep-fake-creation tools have further spurred worries about an "information apocalypse" that destroys "public trust in information and the media."[38] Platforms, in this view, foster a world in which citizens lack a shared reservoir of mutual tolerance and factual beliefs about the world. Such deficiencies are said to render meaningful political debate on social platforms challenging—perhaps even impossible. As a result of these changes in peoples' dispositions, the possibility of democratic life moves out of reach.

These arguments hence assume that democratic life requires the prevalence of certain attitudes and beliefs in order to be durably sustained (an assumption that may or may not be empirically justified). Another way in which these concerns can concretely be understood is to view them in light of the rise of anti-system parties,[39] which are characterized by their limited

34. *See, e.g.*, Helen Margetts, *Rethinking Democracy with Social Media*, 90 THE POL. Q., Jan. 2019, 107, at 107 (assigning blame to social media for "pollution of the democratic environment through fake news, junk science, computational propaganda and aggressive microtargeting and political advertising"; for "creating political filter bubbles"; and for "the rise of populism, . . . the end of democracy and ultimately, the death of democracy.").

35. Jonathan Haidt, *Yes, Social Media Really Is Undermining Democracy*, THE ATLANTIC (July 28, 2022), https://www.theatlantic.com/ideas/archive/2022/07/social-media-harm-facebook-meta-response/670975 [https://perma.cc/7FFV-QRPB].

36. David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts & Jonathan L. Zittrain, *The Science of Fake News: Addressing Fake News Requires a Multidisciplinary Effort*, 359 SCI. 1094, 1094 (2018); *see also* Edson C. Tandoc Jr., *The Facts of Fake News: A Research Review*, SOC. COMPASS, July 25, 2019, at 1, 2 ("[Fake news] is intended to deceive people, and it does so by trying to look like real news."). For examples, see Aziz Z. Huq, *Militant Democracy Comes to the Metaverse?*, 72 EMORY L.J. 1105, 1118–19 (2023). The terms "misinformation" and "disinformation" are also used to describe fake news and its variants. I leave aside questions about how to exactly define and distinguish these terms.

37. *See, e.g.*, Haidt, *supra* note 35; Zeynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 COLO. TECH. L.J. 203, 215 (2015) (criticizing Facebook for its power to "alter the U.S. electoral turnout" through algorithmic manipulation).

38. Mateusz Łabuz & Christopher Nehring, *On the Way to Deep Fake Democracy? Deep Fakes in Election Campaigns in 2023*, 23 EUR. POL. SCI. 454, 457 (2024).

39. Giovanni Capoccia, *Anti-System Parties: A Conceptual Reassessment*, 14 J. THEORETICAL

regard for democratic norms. Platforms might facilitate the growth of such anti-system candidates who disrupt or even undermine democratic norms such as broad trust in the state and in co-citizens. Through this indirect path, platforms have a detrimental effect on democracy's prospects.

There are surprisingly few empirical studies that support the existence of a robust causal connection between social platforms and democratically necessary trust.[40] Yet some evidence for it can be found in the behaviors and beliefs of significant political actors. President Donald Trump, for example, declared in November 2016 that Facebook and Twitter had "helped him win" the 2016 U.S. presidential election.[41] Since 2020, conservative donors such as the Bradley Impact Fund and the Conservative Partnership Fund have contributed millions to Republican-aligned groups combating effects to "take a tougher line against misinformation online."[42] Such significant financial investments by important political actors, beyond merely cheap talk, suggest that social platforms do have predictable partisan effects for candidates and parties that have an arguable anti-systemic orientation.[43]

On the other hand, well-designed empirical studies have cast doubt on the negative, large-"N" effects of social platforms.[44] Four studies are illustrative. A first well-designed randomized experiment, which tested the effect of platform deactivation for several weeks before the 2020 election, found no statistically significant effects of platform exposure on affective polarization, issue polarization, or vote choice.[45] A second random

POL. 9, 10–11 (2002) (offering several different definitions of that term).

40. There is one experiment focused on search ranking that finds political effects, but the experiment is more than a decade old and focuses on how search results are displayed, not on the central issue of platform design today. Robert Epstein & Ronald E. Robertson, *The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections*, 112 PROC. NAT'L ACAD. SCI. E4512, E4518–20 (2015).

41. Rich McCormick, *Donald Trump Says Facebook and Twitter 'Helped Him Win'*, THE VERGE (Nov. 13, 2016, 7:02 PM PST), https://www.theverge.com/2016/11/13/13619148/trump-facebook-twitter-helped-win [https://perma.cc/5MUQ-7R73].

42. Jim Rutenberg & Steven Lee Myers, *How Trump's Allies Are Winning the War Over Disinformation*, N.Y. TIMES, https://www.nytimes.com/2024/03/17/us/politics/trump-disinformation-2024-social-media.html [https://web.archive.org/web/20250401001211/https://www.nytimes.com/2024/03/17/us/politics/trump-disinformation-2024-social-media.html].

43. A mea culpa: in previous work, I was too credulous in respect to claims of platform-related harms. Huq, *supra* note 36, at 1118–19. I should have been more cautious.

44. For a prescient popular argument to that effect, see Gideon Lewis-Kraus, *How Harmful Is Social Media?*, NEW YORKER (June 3, 2022), https://www.newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think [https://perma.cc/7FFV-QRPB].

45. The study found a non-significant pro-Trump effect from Facebook usage but cautioned against treating this finding as generalizable. Hunt Allcott, Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Sandra González-Bailón, Andrew M. Guess, Young Mie Kim, David Lazer, Neil Malhotra, Devra Moehler, Sameer Nair-Desai, Houda Nait El Barj, Brendan Nyhan, Ana Carolina Paixao de

experiment focused on the difference between Facebook's default algorithms and a reverse-chronological feed. Again, the study found no effect on affective polarization, issue polarization, or political knowledge after switching from a network-driven feed to an algorithmically-driven feed, even though the use of a reverse chronological feed increased the amount of "untrustworthy" content seen.[46] This null finding from a study of opting into algorithmic content propagation has been replicated in a separate study of YouTube.[47]

Finally, an empirical inquiry into exposure to fake news found only a very small positive effect on the vote share of populist candidates in European elections.[48] Another study of 1,500 users in each of three countries (France, the United Kingdom, and the United States) identified no correlation between social platform use and more extreme right-wing views; indeed, in the United States, they found a *negative* correlation.[49] The authors concluded that their "findings tend to exonerate the Internet generally and social media in particular, at least with respect to right-wing populism."[50] Finally, a 2017 study found that President Trump erred when he claimed that Twitter and X helped him in the 2016 election; again, that study found a negative correlation between more extreme right-wing views and social platform usage.[51]

Summarizing the available research (including these studies) in a June 2024 issue of *Nature*, a team of respected scholars concluded that "exposure

Queiroz, Jennifer Pan, Jaime Settle, Emily Thorson, Rebekah Tromble, Carlos Velasco Rivera, Benjamin Wittenbrink, Magdalena Wojcieszak, Saam Zahedian, Annie Franco, Chad Kiewiet de Jonge, Natalie Jomini Stroud & Joshua A. Tucker, *The Effects of Facebook and Instagram on the 2020 Election: A Deactivation Experiment*, 121 PROC. NAT'L ACAD. SCI., 1, 8–9 (2024).

46. Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud & Joshua A. Tucker, *How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?*, 381 SCI. 398, 402 (2023).

47. Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild & Duncan J. Watts, *Examining the Consumption of Radical Content on YouTube*, 118 PROC. NAT'L ACAD. SCI., 1, 1 (2021).

48. Michele Cantarella, Nicolò Fraccaroli & Roberto Volpe, *Does Fake News Affect Voting Behaviour?*, RSCH. POL'Y, Jan. 2023, at 1, 2.

49. Shelley Boulianne, Karolina Koc-Michalska & Bruce Bimber, *Right-Wing Populism, Social Media and Echo Chambers in Western Democracies*, 22 NEW MEDIA & SOC'Y 683, 695 (2020).

50. *Id.*

51. Jacob Groshek & Karolina Koc-Michalska, *Helping Populism Win? Social Media Use, Filter Bubbles, and Support for Populist Presidential Candidates in the 2016 US Election Campaign*, 20 INFO., COMMC'N & SOC'Y 1389, 1397 (2017) ("American voters who used social media to actively participate in politics by posting their own thoughts and sharing or commenting on social media were actually more likely to not support Trump as a candidate.").

to misinformation is low as a percentage of people's information diets" and further "the existence of large algorithmic effects on people's information diets and attitudes has not yet been established."[52] The *Nature* team warned that the extent to which social platforms undermine political knowledge depends on the availability of other news sources. Where countries "lack reliable mainstream news outlets," their negative knowledge-related spillovers may be greater.[53] I do not pursue that suggestion here, since it invites a bifurcated analysis that separately considers different national jurisdictions, depending on the robustness of their non-digital media ecosystems. What follows should be taken as parochially relevant to North American and European democracies (at least for now) but not the larger world beyond that.

A second view of social platforms' harms identifies not its spillovers at scale, but rather its effects on certain narrow slices of the population—in particular, those at the tails of the ideological distribution. The intuition here is that engagement with social platforms may not change the dispositions or beliefs of most people, but there is a small subset of individuals who adopt dramatically divergent beliefs (and even behaviors) as consequences of their platform use. "Tail effects" of this sort may not be significant for democratic life under some circumstances, but of particular relevance, there is some evidence of increased support for political violence among Americans.[54] Extremism at the tails in *this* context and about *this* sentiment may have profound consequences. At a moment when President Trump has (twice) faced near-assassination during the 2024 presidential election cycle, and considering how his supporters previously precipitated a deadly confrontation at a 2021 Joint Session of Congress meant to count Electoral College votes, it seems prudent to reckon with the risk that radicalized individuals—even if few in number—may be able to inflict disproportionate harms on institutions that are necessary for core democratic political processes.

52. Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson & Duncan J. Watts, *Misunderstanding the Harms of Online Misinformation*, 630 NATURE 45, 47–48 (2024); *accord* Sacha Altay, Manon Berriche & Alberto Acerbi, *Misinformation on Misinformation: Conceptual and Methodological Challenges*, SOC. MEDIA + SOC'Y, Jan.–Mar. 2023, at 1, 3 ("Misinformation receives little online attention compared to reliable news, and, in turn, reliable news receives little online attention compared to everything else that people do.").

53. Budak et al., *supra* note 52, at 49.

54. At least some surveys suggest rising levels of positive attitudes to violence. *See* Ashley Lopez, *More Americans Say They Support Political Violence Ahead of the 2024 Election*, NPR, https://www.npr.org/2023/10/25/1208373493/political-violence-democracy-2024-presidential-election-extremism [https://perma.cc/ZM4L-BRRV]. For other findings exhibiting a concentration of such support at the rightward end of the political spectrum, see Miles T. Armaly & Adam M. Enders, *Who Supports Political Violence?*, 22 PERSP. ON POL. 427, 440 (2024).

This more narrowly gauged claim stands on firmer empirical ground than the critiques of social platforms' large-"N" effects discussed above. A 2024 study of fake news' circulation on Twitter found that 0.3 percent of users account for four-fifths of its fake news volume.[55] These "supersharers," who tended to be older, female, and Republican, in turn reached a "sizable 5.2% of registered voters on the platform."[56] Note that this is not necessarily the population one would expect to engage in political violence. A different study published around the same time also found "asymmetric . . . political news segregation" with "far more homogenously conservative domains and URLs circulating on Facebook" and "a far larger share" of fake news on the political right.[57]

Such findings are consistent with wider-angle studies of partisan polarization, which find different microfoundations on the political left and right.[58] The *Nature* team mentioned above hence concluded that exposure to misinformation is "concentrated among a small minority."[59] Those who consume false or otherwise potentially harmful content are already attuned to such information and actively seek such content out.[60] Platforms, however, do not release "tail exposure metrics" that could help quantify the risk of harm from such online interactions.[61] As a result, it is hard to know how serious the problem may be.

What of the concern that social platforms conduce to "filter bubbles" that constrain the range of information sources users can access in problematic ways?[62] Once again, the evidence is at best inconclusive. A 2016 study found that social homogeneity of users predicted the emergence of

55. Sahar Baribi-Bartov, Briony Swire-Thompson & Nir Grinberg, *Supersharers of Fake News on Twitter*, 384 SCI. 979, 980 (2024).

56. *Id.* at 979.

57. Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud & Joshua A. Tucker, *Asymmetric Ideological Segregation in Exposure to Political News on Facebook*, 381 SCI. 392, 397 (2023).

58. Craig M. Rawlings, *Becoming an Ideologue: Social Sorting and the Microfoundations of Polarization*, 9 SOCIO. SCI. 313, 337 (2022).

59. Budak et al., *supra* note 52, at 48.

60. *Id.*

61. *Id.* at 50; *see also* Vivian Ferrillo, *r/The_Donald Had a Forum: How Socialization in Far-Right Social Media Communities Shapes Identity and Spreads Extreme Rhetoric*, 52 AM. POL. RSCH. 432, 443 (2024) (finding that users who engage often with a far-right community also use far-right vocabulary more frequently in other spaces on their platform, contributing to the spread and normalization of far-right rhetoric).

62. For an influential treatment of the topic, see generally ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2012).

echo chambers characterized by asymmetrical patterns of news sharing.[63] At the same time, the study offered no empirical evidence about the extent or effects of filter bubbles "in the wild," so to speak. A 2021 review identified divergent results in studies surveying human users of social platforms or digital trace data; yet, it identified only a handful of studies substantiating the concern.[64] A 2022 meta-study found that "most people have relatively diverse media diets," and only "small minorities, often only a few percent, exclusively get news from partisan sources."[65] Again, the empirical foundations of the normative worry here seem shaky.

Even if the evidence of filter bubbles existing was more robust, filter bubbles' substantiated existence would not necessarily be cause for concern. Concern about filter bubbles focuses on the asymmetric character of the information voters consume; this then assumes that there is a counterfactual condition under which the voter might receive a "balanced" diet of information. But what does it mean to say that a person's news inputs are balanced or symmetrical? Does it require equal shares of data that support Republican and Democratic talking points? What if one of those parties is more likely than the other to lean on false empirical claims? Should a balanced informational diet reflect or discount for such a lean? How are the problems of misinformation or distorted information to be addressed? Is it part of a balanced informational diet to receive a certain amount of "fake news"? These questions admit of no easy answers. Rather, they suggest that the concern with filter bubbles trades on a notion of balance that is hard to cash out in practice without difficult anterior ideological and political choices.

In brief, the available empirics suggest that consequentialist critiques of social platforms are better focused on tail effects instead of the way platform engagement changes the median user or the mass of users. It is also worth underscoring a point that is somewhat obscured by the bottom-line results of these studies but implicit in what I have just set out. That is, the tail effects of social platforms arise from a complex and unpredictable mesh of interactions between technical design decisions and users' decisions. The external political environment hence shapes platforms' spillover effects, and

63. Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley & Walter Quattrociocchi, *The Spreading of Misinformation Online*, 113 PROC. NAT'L ACAD. SCI. 554, 558 (2016).

64. Ludovic Terren & Rosa Borge, *Echo Chambers on Social Media: A Systematic Review of the Literature*, 9 REV. COMMC'N RSCH. 99, 110 (2021) (reviewing fifty-five studies and finding only five yielding no evidence of echo chambers).

65. AMY ROSS ARGUEDAS, CRAIG T. ROBERTSON, RICHARD FLETCHER & RASMUS K. NIELSEN, ECHO CHAMBERS, FILTER BUBBLES, AND POLARISATION: A LITERATURE REVIEW 4 (2022), available at https://ora.ox.ac.uk/objects/uuid:6e357e97-7b16-450a-a827-a92c93729a08.

when that environment is more polarized and more prone to panics or even violence, it seems likely that the tail risks of social platforms would correspondingly rise. When, by contrast, there are a plethora of reliable non-digital sources which are accurate and easily accessible, the threat to democratic life from social platforms may well be far less acute.

C. DEONTIC CRITIQUES OF SOCIAL PLATFORMS

Critiques of social platforms do not need to rest on evidence of their consequences. It is also possible to pick out features of the relationship between platforms and users as morally problematic even in the absence of any harm arising. Two particular strands of such "deontic" critique can be traced in existing literature.

First, social platforms (among other entities) gather data about their users and then use that data to target advertisements to those same users. For many, this circular pattern of data extraction and deployment constitutes a morally problematic exploitation. Such exploitation occurs when "one party to an ostensibly voluntary agreement intentionally takes advantage of a relevant and significant asymmetry of knowledge, power, or resources" to offer otherwise unacceptable contracting terms.[66]

Shoshana Zuboff, who is perhaps the leading expositor of this view, argues that platforms have "scraped, torn, and taken for another century's market project" the very stuff of "human nature."[67] She condemns the "rendition" and "dispossession of human experience" through "datafication."[68] Zuboff's critique of platform exploitation is nested in a broader set of concerns about the presently hegemonic form of "informational" or "financial" capitalism. Reviewing Zuboff's book, Amy Kapczynski thus asserts that "informational capitalism brings a threat not merely to our individual subjectivities but to our ability to self-govern."[69] Similarly, danah boyd characterizes private firms' use of digital power as a malign manifestation of "late-stage capitalism . . . driven by financialization."[70] And as Katharina Pistor puts it, "[t]he real threat that emanates from Big Tech using big data is not just market dominance . . . [but] the power to transform free contracting and markets

66. Claire Benn & Seth Lazar, *What's Wrong with Automated Influence*, 52 CANADIAN J. PHIL. 125, 135 (2022).

67. SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 94 (2019).

68. *Id.* at 233–34.

69. Amy Kapczynski, *The Law of Informational Capitalism*, 129 YALE L.J. 1460, 1467 (2020).

70. danah boyd, *The Structuring Work of Algorithms*, 152 DÆDALUS 236, 238 (2023).

into a controlled space that gives a huge advantage to sellers over buyers."[71] The structure of financial or quasi-financial transactions on social platforms, in this view, conduces systemically to users' exploitation.

In an earlier piece, I have expressed sharp skepticism elsewhere about the empirical and normative arguments offered by Zuboff and Kapczynski.[72] Their concerns about exploitation seem to trade on imprecise and potentially misleading analogies to more familiar and normatively troubling forms of economic exploitation, despite meaningful differences in structure and immediate effect. Indeed, both analogies fail to take those differences seriously. More generally, their arguments borrow a suite of concerns associated with the larger structures of economic life labeled "neoliberalism," which have developed since the 1970s. Such critiques of neoliberalism, however, concern aspects of economic life that have little to do with social platforms (for example, deregulation and financialization). One can have neoliberalism with or without social platforms. I see little analytic gain in combining these very different lines of argument respecting quite distinct targets, and I see no reason to invite confusion by mushing together distinct phenomena to achieve guilt-by-association more generally.

Second, concern about exploitation overlaps with a distinct worry about non-domination. Claire Benn and Seth Lazar capture this possibility in their argument that social platforms might compromise an intrinsic, non-instrumental "value of living in societies that are free and equal."[73] They argue that the public is necessarily ignorant about the "tech companies' control of the means of prediction" and so have "no viable way of legitimating these new power relations."[74] But the empirical premise of this argument—widespread public ignorance about predictive tools—seems shaky: As the empirical studies of fake news and political distortion show, there is publicly available knowledge about many salient effects of social platforms. To the extent that the public misconstrues those effects, Benn and Lazar's argument likely overestimates their magnitude.[75] I hardly think these critiques are secret.

Still, I think Benn and Lazar are on to something useful when they identify the fact of corporate control as a morally salient one. Social

71. Katharina Pistor, *Rule by Data: The End of Markets?*, 83 LAW & CONTEMP. PROBS. 101, 117 (2020); *accord* Julie E. Cohen, *Law for the Platform Economy*, 51 U.C. DAVIS L. REV. 133, 145–48 (2017).

72. Mariano-Florentino Cuéllar & Aziz Z. Huq, *Economies of Surveillance,* 133 HARV. L. REV. 1280, 1298 (2020).

73. Benn & Lazar, *supra* note 66, at 133.

74. *Id.* at 137.

75. *See supra* notes 35 and 37 for examples of such overestimation.

platforms stand in an asymmetrical relation to the general public because of (1) knowledge asymmetries enabled by the corporate form; (2) collective action problems implicit in the one-to-many relation of firms to consumers; and (3) ideological effects (for example, false beliefs in the necessity of unregulated digital markets for economic growth). As a consequence of these dynamics, social platforms exercise a certain kind of unilateral power over the public. Such power might be especially worrying if it is concentrated in the hands of a limited number of people—and if these people have close connections to those in high state office (with the Musk/Trump relationship offering an obvious, highly salient example). This slate of worries comes sharply into play whenever platforms comprise an important part of the democratic public sphere. Under these conditions, Benn and Lazar point out that platforms ought not to merely prevent negative consequences for democratic politics; they must also ensure "that content promotion is regulated by epistemic ideals."[76] This entails, in their view, a measure of "epistemic paternalism."[77] It rests on platforms' unilateral, and effectively unconstrained, judgments about interface and algorithmic design.

This deontic argument can also be stated in the terms of Philip Pettit's influential theory of republican freedom. On Pettit's account, an individual wields dominating power over another if the former has the capacity to interfere in certain choices of the latter on an arbitrary basis.[78] Pettit asserts that this arbitrariness condition is satisfied if an agent's actions are subject only to the *arbitrium*—the will or judgment—of the agent, and when the interfering agent is not "forced to track the interests and ideas of the person suffering the interference."[79] For example, a person ranked by law as a slave is just as unfree even if their master always acts with their interests in mind. Even when an arbitrary legal relationship is exercised in a beneficent fashion with the interest of the weaker party in mind, Pettit suggests that there is a displacement of the subject's "involvement, leaving [them] subject to relatively predictable and perhaps even beneficial forms of power that nevertheless 'stifle' and 'stultify.' "[80]

Yasmin Dawood has fruitfully deployed Pettit's framework for thinking about the abuse of public power in democratic contexts.[81] Her conceptual framing, moreover, could be extended to private actors such as social

76. Benn and Lazar, *supra* note 66, at 144.

77. *Id.*

78. PHILIP PETTIT, REPUBLICANISM: A THEORY OF FREEDOM AND GOVERNMENT 52 (1997).

79. *Id.* at 55.

80. Patchen Markell, *The Insufficiency of Non-Domination*, 36 POL. THEORY 9, 12 (2008). To be clear, Markell here is criticizing and extending Pettit's account.

81. Yasmin Dawood, *The Antidomination Model and the Judicial Oversight of Democracy*, 96 GEO. L.J. 1411, 1431 (2008).

platforms without too much difficulty. For instance, one might view the exercise of extensive control over the informational environment online as a species of domination, whether or not it was exercised in a malign or a paternalistic direction. That idea might be rendered more precise by drawing on work by Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner that defines the "performative power" of an algorithmic instrument in terms of "how much participants change in response to actions by the platform, such as updating a predictive model" as a numerical parameter.[82] This concept of "performative power" usefully captures the way that platforms "steer" populations.[83] As such, it offers a way of understanding and measuring "domination" in social platforms more precisely.

In setting out these two kinds of deontic critiques of social platforms, I thus suggest that there are plausible grounds for worry about social platforms, even absent robust empirical findings of spillovers onto users' beliefs and dispositions. I recognize that both the exploitation and the domination critiques rest on further moral premises, which I have neither spelled out in full nor tried to substantiate. But I spell out both deontic arguments here to show readers the minimally plausible non-consequentialist grounds for concern about the structure and operation of social platforms and to allow readers to make their own judgments.

D. MAKING A BETTER CASE AGAINST SOCIAL PLATFORMS

Social platforms have become scapegoats of sorts for many of the ills that democratic polities are now experiencing. But the available evidence suggests that many of these critiques miss the mark. For many people, platforms simply do not play a very large or dominant epistemic role (although this may well change in the near future). They also seem not to have the polarizing and epistemically distorting effects many bemoan.

That is not to say, however, that there is no reason for concern. Consequentialist worries about the behavior of users on the tails of the ideological distribution, as well as deontic worries about exploitation or domination, point toward the need for reforming measures. Of course, these arguments might not all point in the same direction in terms of practical change. But reforms that render platforms more responsive and responsible to epistemically grounded truths and the interests of all their users (as well as interests of the general public at large) are plausibly understood to respond to all the salient critiques discussed above.

82. Moritz Hardt, Meena Jagadeesan & Celestine Mendler-Dünner, *Performative Power*, 2022 NIPS '22: PROC. OF THE 36TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 2.

83. *Id*. at 5–6.

II. ISLANDS OF INTEGRITY—REAL AND DIGITAL EXAMPLES

One way of thinking about how existing social platforms might be reformed is to identify an aspirational end-state, or a model, of how they might ideally work. With an understanding of the best version of a social platform in view, it may be easier to evaluate extant reform strategies and to propose new ones. This inquiry might proceed at the retail level—focusing on what an "ideal" or a "better" platform might look like—or at a general level—asking how the digital ecosystem overall should be designed. With the first of these paths in mind, I introduce in this Part a conceptual framework for thinking about "islands of integrity" developed in the sociological and political science studies of development. While that literature has not yielded any simple or single formula for reaching that aspiration, it still offers a useful lens for starting to think about well-functioning social platforms. Or so I hope to show in what follows.

A. BUILDING ISLANDS OF INTEGRITY IN THE REAL WORLD

In recent decades, concern about the legality and the quality of governance has shaped the agenda of international development bodies such as the World Bank.[84] One of the strategies identified to enhance the quality of public administration centers the idea of "islands of integrity" or "pockets of effectiveness" in sociopolitical environments that are "otherwise dominated by patrimonialism, corruption, and bureaucratic dysfunction."[85] An island of integrity has been defined as an entity or unit (generally of government) that is "reasonably effective in carrying out [its] functions and in serving some conception of the public good, despite operating in an environment in which most agencies are ineffective and subject to serious predation . . . ."[86] The normative intuition is that it is possible to seed islands of integrity, despite pervasive corruption, as a starting point for more large-scale reforms.

There are by now a wide variety of case studies on islands of integrity. Monica Prasad, for example, points to the recently stood-up Indian Institutes of Technology ("IITs"), an archipelago of meritocratic technology-focused colleges across the subcontinent, as an instance where an educational

84. AZIZ Z. HUQ, THE RULE OF LAW: A VERY SHORT INTRODUCTION 75–78 (2024).

85. Prasad, *supra* note 15, at 376.

86. David K. Leonard, *'Pockets' of Effective Agencies in Weak Governance States: Where Are They Likely and Why Does It Matter?*, 30 PUB. ADMIN. & DEV. 91, 91 (2010); *see also* Michael Roll, *The State That Works: A 'Pockets of Effectiveness' Perspective on Nigeria and Beyond*, *in* STATES AT WORK: DYNAMICS OF AFRICAN BUREAUCRACIES 365, 367 (Thomas Bierschenk & Jean-Pierre Olivier de Sardan eds., 2014) ("A pocket of effectiveness (PoE) is defined as a public organisation that provides public services relatively effectively despite operating in an environment, in which public service delivery is the exception rather than the norm.").

mission is successfully pursued against "a context of patrimonialism and corruption."[87] IITs' mission is preserved and protected from distortion through the use of selection strategies of "meritocratic decoupling" that sort both students and teachers based on academic merit, alongside efforts to show how the institution benefited those who were excluded.[88]

In a different case study, Eliška Drápalová and Fabrizio Di Mascio identify a pair of municipalities in Spain as "islands of integrity."[89] They contend that the key move in creating them was the fashioning of a "fiduciary relationship between mayors and city managers," which allowed for the development of a bureaucratic structure shaped by professional (rather than patrimonial) norms.[90] City managers, they find, offer "accountability and responsiveness" to elected leaders without compromising the integrity of service-oriented institutions.[91] Similarly, Michael Roll maps the emergence in Nigeria of well-run agencies managing food and drug regulation on the one hand, and human trafficking on the other, to demonstrate that islands of integrity can emerge even under very difficult circumstances given the right leadership.[92]

Most, but not all, of these case studies on islands of integrity concern real-world public administration, often at a local level.[93] The generalizations drawn by the literature are concededly fragile: The heterogeneity of cultural, political, and institutional context makes inference instable, at least at a useful level of granularity.[94] Still, a couple of regularities do tentatively emerge from a review of the available case studies in the development literature.

Crudely stated, the "islands of integrity" literature underscores the importance of institutional *means* and leadership *motives* for resisting patrimonial or corrupt political cultures. First, an island of integrity needs to internalize control over its own workings in order to "create a culture of meritocracy and commitment to the organization's mission."[95] Underpinning this culture, it seems, must be a clear understanding of the public goods that

87. Prasad, *supra* note 15, at 380.
88. *Id.* at 382–83.
89. Drápalová & Di Mascio, *supra* note 15, at 128.
90. *Id.* at 129–30, 135.
91. *Id.* at 135.
92. Roll, *supra* note 86, at 370–73.
93. One article applies the concept to public broadcasters in developing countries, but does not do so with enough detail to be useful. Cherian George, *Islands of Integrity in an Ocean of Commercial Compromises*, 45 MEDIA ASIA 1, 1–2 (2018).
94. Leonard compiles a number of general lessons, but these are pitched at a very high level of abstraction. Leonard, *supra* note 86, at 93.
95. Prasad, *supra* note 15, at 376.

the agency or body is supposed to produce. The truism that leadership is key seems to hold particularly strongly.[96] Autonomy over personnel choice is also crucial in order to maintain that culture.[97]

Second, there is a consistent institutional need for the creation of tools to resist demands from powerful external actors who try to capture a body for their immediate political or economic gains, which are unrelated to the public-regarding goals of the institution.[98] Tools by which to mitigate such threats to institutional autonomy vary. Indian universities, Prasad found, tout the local jobs they create in cleaning and maintenance—even as they maintain the separation of student and faculty selection from local pressures—as a way of deflecting local politicos.[99] Spanish city managers, Drápalová and Di Mascio explain, promise improvements in top-line municipal services to mayors who threaten their autonomy.[100] In effect, reform is purchased in both cases by strategic payoffs to those who threaten its progress.

Just as it is important to work out how to build public-regarding institutional spaces in the real world, so too is it important to identify how to create such spaces in the virtual, digitally mediated world. Just as the bodies in India, Spain, and Nigeria need to have motive and means to keep the corroding forces of public sphere at bay, so too does a social platform that strives to be an island of integrity need leadership, internal culture, and means to create a non-exploitative, non-dominating structure while managing tail risk better than existing platforms. Taken as metaphor, therefore, "islands of integrity" offer a template for the desirable end goal of social platform reform as well as some modest clues about how to get there. Still, it is important not to make too much of this metaphor. The claim that the "islands of integrity" literature can be an inspiration for social platform reform is, at bottom, an argument from analogy, and one that needs to be tested carefully through the application of that analogy.

B. DIGITAL ISLANDS OF INTEGRITY: TWO EXAMPLES

The aforementioned analogy gains force when one realizes that there are already examples of digital islands of integrity online. The two most prominent examples are Wikipedia and the British Broadcasting Company ("BBC"). To be clear, neither is a quintessential social platform as I have

96. Leonard, *supra* note 86, at 94 (noting the importance of "leadership, personnel management, resource mobilisation and adaptability").

97. Roll, *supra* note 86, at 379.

98. *Id.* at 377–78 (noting the role of tools for "political management").

99. Prasad, *supra* note 15, at 385.

100. Drápalová and Di Mascio, *supra* note 15, at 135.

used that term here. Nor do they operate at the same scale as X or Instagram. But I offer a brief discussion of both by way of proof of concept.

Wikipedia emerged from the wreckage of an attempted for-profit online encyclopedia called Nupedia.[101] The latter's assets (for example, domain names, copyrights, and servers) were subsequently placed in an independent, charitable organization, the Wikimedia Foundation ("WMF").[102] At first, corporate governance "emerged" organically from the efforts of those building the site, rather than being imposed from above.[103] A group of founders then "transformed their charismatic community into a bureaucratic structure" in which "power was diffused and distributed" across "a sprawling bureaucracy with a wide range of formal positions" in response to the perceived mission-related needs of the organization.[104] The organization's trajectory has also been characterized by moments of greater centralization. For example, in the early 2010s, Wikipedia's CEO led an effort to be "more inclusive and more open," somewhat to the chagrin of the then-contributors.[105] That is, Wikipedia's governance history centers on a choice of corporate form that insulates leadership from external profit-related pressures, a selection of strong leadership, and an exercise of leadership to broaden and empower the organization's constituencies (potentially mitigating criticism of the organization) to generate a certain kind of "corporate culture."[106]

Even more directly relevant is the web presence of the BBC. The BBC produces thousands of new pieces of content each day for dissemination over a range of sites, such as BBC News, BBC Sport, BBC Sounds, BBC iPlayer, and World Service.[107] The corporation's charter defines its mission as serving all audiences by providing "impartial, high-quality and distinctive output and services which inform, educate and entertain."[108] Like Wikipedia, the BBC is organized into a corporate form that is relatively impermeable by

101. Emiel Rijshouwer, Justus Uitermark & Willem de Koster, *Wikipedia: A Self-Organizing Bureaucracy*, 26 INFO., COMMC'N & SOC'Y 1285, 1291 (2023).

102. *Id.* at 1293.

103. *Id.* at 1298 (explaining that "bureaucratization emerges from interactions among constituents" of Wikipedia).

104. *Id.* at 1294.

105. *Id.* at 1296.

106. *Cf.* Pasquale Gagliardi, *The Creation and Change of Organizational Cultures: A Conceptual Framework*, 7 ORGANIZATIONAL STUD. 117, 121–26 (1986) (exploring the meaning of the term "organizational value" and culture).

107. Alessandro Piscopo, Anna McGovern, Lianne Kerlin, North Kuras, James Fletcher, Calum Wiggins & Megan Stamper, *Recommenders with Values: Developing Recommendation Engines in a Public Service Organization*, KNIGHT FIRST AMEND. INST. (Feb. 5, 2024), https://knightcolumbia.org/content/recommenders-with-values-developing-recommendation-engines-in-a-public-service-organization [https://perma.cc/APX5-T9T2].

108. *Id.*

law to commercial pressures. To advance its charter goals, the BBC uses machine-learning recommender algorithms created by multi-disciplinary teams of data scientists, editors, and product managers.[109] Once a recommender system has been built,[110] editorial staff can offer "continuous feedback" on the design and operation of recommendatory systems to identify legal compliance questions and to ensure "BBC values" are advanced.[111]

Available accounts of this process—while perhaps a touch self-serving because they are written by insiders—suggest that the organization strives to cultivate a distinctive cultural identity. It then leverages that identity as a means of advancing its values via algorithmic design. Specifically, an important part of this recommender design process focuses on empowering users to make their own choices and to avoid undesirable (from the service's perspective) behaviors. The BBC's recommender tools are designed to permit personalization, albeit only to the extent that doing so can "coexist with the BBC's mission and public service purposes."[112] An insider informant speaking anonymously reported that the BBC understands itself as "as 'morally obliged' to provide their users with the possibility of tweaking their recommendations."[113] In the same study, the employee of an unnamed European public broadcaster that managed a recommender system reported that their system proactively identified "users who consume narrow and one-sided media content and recommend[ed to] them more diverse content."[114] That is, the system was designed to anticipate and mitigate, to an extent, the possibility of extremism at the tails of the user distribution, while also preserving users' influence over the content of their feeds. This is in stark contrast to systems that are designed to maximize engagement under conditions in which predictability entails driving users to more extreme (and even dangerous) content.

I do not want to strain the parallels between the "islands of integrity" literature and these digital examples too much. Both of the latter, nevertheless, point to ways in which the means and the motives to sustain an "island of integrity" can be imagined in an online world. Both services are, for example, explicitly oriented toward a public service mission in terms of leadership. They both also opted for corporate forms that allow for some

109. *Id.*

110. *Id.* Public service broadcasters such as the BBC cannot rely on "off-the-shelf" recommenders because they optimize for very different goals. Jockum Hildén, *The Public Service Approach to Recommender Systems: Filtering to Cultivate*, 23 TELEVISION & NEW MEDIA 777, 787 (2022).

111. Piscopo et al., *supra* note 107.

112. *Id.*

113. Hildén, *supra* note 110, at 786.

114. *Id.* at 788.

protection against potentially compromising market forces. Both have systems in place to preserve and transit a valued internal culture, while buffering themselves somewhat against the risks of distorting external or internal pressure. Finally, both seem to have successfully cultivated persisting cultures of service to public-service goals by hard-wiring their cultures into bureaucratic structures or, alternatively, algorithmic designs.

III. THE GOVERNANCE OF SOCIAL PLATFORMS: ASPIRING TO BUILD ISLANDS OF ALGORITHMIC INTEGRITY

With the general idea of "islands of integrity" in hand, along with the specific proofs of concept described in Section II.B, it is possible to ask how certain social platforms might be reformed with an ideal of islands of algorithmic integrity in mind. That is, how might we move toward alternative platform designs and operations that address the normative concerns outlined in Part II? What kind of private governance might be imagined that mitigates exploitation and domination concerns, while addressing the tail risk of extremism as best as we can? Could legal regulation play a role? Again, it would be a mistake to frame these questions as *mechanical* applications of the "islands of integrity" literature. It is better to think of them as falling out of the same institutional design goal.

I approach this inquiry in two stages. I first begin by critiquing leading regulatory strategies observed in the American states and the European Union from an "islands-of-algorithmic-integrity" standpoint. At bottom, these critiques draw out ways in which those regulatory strategies take social platforms as potential sources of harm, largely without an account of the positive role platforms could play. Second, I draw together a number of possible tactics for public or private actors to help build islands of algorithmic integrity. My positive accounting here is concededly incomplete. My hope, however, is that this effort serves as initial evidence of the fruitfulness of an approach oriented toward the aspiration of islands of algorithmic integrity.

A. THE LIMITS OF EXISTING PLATFORM REGULATION REGIMES

Since 2020, social platforms have become an object of regulatory attention on both sides of the Atlantic. Three main regulatory strategies can be observed. These take the form of new state regulations purportedly targeting "censorship,"[115] fresh efforts to extend common law tort liabilities

115. Mary Ellen Klas, *DeSantis Proposal Would Protect Candidates Like Trump from Being Banned on Social Media*, MIA. HERALD, https://www.miamiherald.com/news/politics-government/state-politics/article248952689.html [https://web.archive.org/web/20221017063802/https://www.miamiherald.com/news/politics-government/state-politics/article248952689.html] (quoting Florida governor Ron

to social platforms, and a risk-based regulatory regime that has been promulgated by the European Union. Broadly speaking, all such legal intervention is premised on concern about platforms' society-wide effects on listeners, although deontic concerns may play a role too. The tools seized for those tasks, however, have been inadequate. Their shortfall can be traced to the way in which they focus exclusively on platform harms (missing the importance of benefits), misconstrue those harms, and then fail to incentivize the formation of platforms with the means and the motive to mitigate documented harms while resisting exploitation or domination.

1. Regulating Ex Ante for Harms

The 2022 Digital Services Act ("DSA") offers a first model of ex ante platform regulation. In important part, it trains on the potential for harms by recommender systems without any account of their positive effects. It contains a suite of new legal obligations: Article 25, for example, prohibits any digital platform design that "deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions."[116] Article 38 provides a right to opt out of non-personalized algorithms.[117] Articles 14 and 26 set out some disclosure obligations on certain companies.[118] Most importantly, for present purposes, Article 34 of the DSA requires "very large online platforms and . . . online search engines" to conduct an annual assessment of any systemic risks stemming from the design or functioning of their service, including negative effects on civic discourse, electoral processes, or fundamental rights.[119]

At first blush, the DSA seems oriented toward the creation of islands of algorithmic integrity. But there are reasons for being skeptical of conceptualizing the project this way. To begin with, the Article 38 opt-out is unlikely to be exercised by those "supersharers" at the tails of the ideological distribution who are most responsible for the diffusion of fake news.[120] Self-help remedies never avail those already fixated on harming themselves and others. Moreover, Article 34 risk assessments impose no clear affirmative command to build epistemically robust speech environments.[121] In effect, the

DeSantis).

116. Digital Services Act, *supra* note 13, at art. 25 § (1).

117. *Id.* at art. 38 (mandating "at least one option for each of their recommender systems which is not based on profiling as defined in Article 4, point (4), of Regulation (EU) 2016/679 ").

118. *Id.* at art. 14 § (1) and art. 26 § (1)(d).

119. *Id.* at art. 34. For a close reading of Article 34, see Neil Netanel, *Applying Militant Democracy to Defend Against Social Media Harms*, 45 CARDOZO L. REV. 489, 566 (2023).

120. Baribi-Bartov et al., *supra* note 55, at 979.

121. *But see* Netanel, *supra* note 119, at 566–67 (proposing that platforms be required to make "recommender system modifications to improve the prominence of authoritative information, including

act offers no clear account of how social platforms could or should enable democratic life. Even more problematic, the DSA ultimately leans on platforms themselves to accurately document and remedy their own flaws. It does not seem excessively cynical to predict that profit-oriented companies will not be falling over themselves to flag the negative externalities of their own products in publicly available documents and flagellate themselves over how to remedy them. The DSA, in short, is promising as theory. But it may fall substantially short in practice.

2. Regulating Ex Ante for Balance

Both Florida and Texas have enacted statutes intended to limit platforms' abilities to "deplatform" a person because of their violation of terms of service.[122] The Florida statute, for example, prohibits platforms from "willfully deplatform[ing] a candidate for office who is known by the social media platform to be a candidate, beginning on the date of qualification and ending on the date of the election or the date the candidate ceases to be a candidate."[123] In its July 2024 decision in *Moody v. NetChoice*, the U.S. Supreme Court cast doubt on the constitutionality of such measures.[124] While litigation is ongoing as this essay goes to press, it seems likely that the deplatforming elements of both statutes will not survive.

Relying on familiar doctrinal tools from the First Amendment toolkit, a majority of the *Moody* Court reached two conclusions that are relevant here. First, Justice Elena Kagan's majority opinion explained that when an entity "provide[s] a forum for someone else's views" and is thereby "engaged in its own expressive activity, which the mandated access would alter or disrupt," a First Amendment interest is implicated.[125] Second, the Court held that the government has no constitutionally cognizable interest "in improving, or better balancing, the marketplace of ideas."[126] This anti-distortion argument is familiar from the campaign finance context.[127] There, however, the argument is deployed generally by conservative justices to

news media content that independent third parties have identified as trustworthy"). Netanel, however, is proposing in this passage an extension of Article 34 rather than offering a gloss on it, so far as I can tell.

122. Florida defines "deplatform" as "the action or practice by a social media platform to permanently delete or ban a user or to temporarily delete or ban a user from the social media platform for more than 14 days." FLA. STAT. § 501.2041(1)(c) (2021). Texas's law has a similar provision. *See* H.B. 20, 87th Leg., Reg. Sess. (Tex. 2021) (prohibiting social media platforms from censoring users or a user's expressions based on the viewpoint expressed in the content).

123. FLA. STAT. § 106.072(2) (2021).

124. Moody v. NetChoice, LLC, 603 U.S. 707 (2024).

125. *Id.* at 728.

126. *Id.* at 732.

127. *See, e.g.*, Citizens United v. FEC, 558 U.S. 310, 340–41 (2010) ("By taking the right to speak from some and giving it to others, the Government deprives the disadvantaged person or class of the right to use speech to strive to establish worth, standing, and respect for the speaker's voice.").

resist governmental efforts to advance an equality interest in political speech given its "dangerous[] and unacceptable" effects.[128] In the Florida and Texas cases, by contrast, the argument was listed against efforts by Republican state governments to enforce their understanding of balance on the platform-based speech. Such ideological valence thus flipped from campaign finance to platform regulation.

Independent of these familiar constitutional logics, there are more empirically grounded reasons to conclude that Florida's and Texas's efforts to mitigate platforms' curatorial capacity are likely to undermine, rather than promote, the emergence of islands of algorithmic integrity. These reasons run parallel to Justice Kagan's reasoning, but are distinctive in character.

The first reason is banal and empirical. The available research suggests that conservative voices in the United States are asymmetrically responsible for the dissemination of fake news.[129] To the extent that Florida and Texas leaned on a conception of "balance" in the speech environment, they did so by culpably ignoring the platforms' interest in a generally reliable and trustworthy news environment. Enforcement of the Florida and Texas laws, to the contrary, seems likely to lead (all else being equal) to a decline in the quality of those platforms. That is to say, by a sort of Gresham's law for political speech, the increasing proportion of misleading speech on a platform will tend to drive out those concerned with truthfulness. Such an effect creates a vicious circle of sorts, one that is absent from the campaign finance context.

This argument might be supplemented by a further observation. As I show below, there are a number of fairly obvious affirmative measures that private and public actors can take if they are truly concerned with the creation of islands of algorithmic integrity.[130] If we see a government failing to take these needful steps while affirmatively adopting counterproductive measures, there is some reason to doubt the integrity of its claim to be acting in the public interest. The islands of algorithmic integrity frame can be put to work here as a lens through which one may understand the gap between a state's professed interests and its actual ambitions.[131] If, as Justice Kagan

128. *Id.* at 351.

129. Baribi-Bartov et al., *supra* note 55, at 979 ("Supersharers had a significant overrepresentation of women, older adults, and registered Republicans."); González-Bailón et al., *supra* note 57, at 397 ("We also observe on the right a far larger share of the content labeled as false by Meta's 3PFC."). There is more to be said about rhetorical use of "balance" claims in law and politics, and its dynamic effects upon the tendency of people to go to extremes.

130. *See infra* Part III.B.

131. *Cf.* Geoffrey R. Stone, *Free Speech in the Twenty-First Century: Ten Lessons from the Twentieth Century*, 36 PEPP. L. REV. 273, 277 (2009) (noting that "government officials will often defend their restrictions of speech on grounds quite different from their real motivations for the suppression,

once suggested in her academic role, the First Amendment doctrine is best understood as "a series of tools to flush out illicit motives and to invalidate actions infected with them" and a "kind of motive-hunting,"[132] then the failure to pick low-hanging fruit while making elaborated and far-fetched claims about one's integrity-related aims is a telling one. To the extent that it identifies some of those low-hanging fruit, the islands of algorithmic integrity grafts on comfortably to advance those goals.

A second reason to be skeptical of measures such as Florida's and Texas's is conceptual in character: balance-promoting measures of their ilk help themselves to the assumption that there is a neutral baseline that has been disturbed by a platform's algorithm. But "the most common choice of baseline fundamentally depends on the state of some components of the system," and assumes away the effect of past bias and amplification.[133] Accordingly, the Florida and Texas laws' presupposition of a neutral baseline of undistorted speech is misplaced; it is better to instead focus on the structural qualities associated with islands of integrity. Where a government asserts an interest in "neutrality" or "fairness" in the context of social platforms, its arguments should be viewed as pro tanto dubious since it is striving to return to a status quo that, for technological reasons, is imaginary. A version of this baseline difficulty arises in the campaign finance context, albeit for different reasons.[134] It also lacks the sociotechnical foundation that is present in the platform context.

3. Tort Liability for Harmful Algorithmic Design

The Texas and Florida statutes impose ex ante controls on social platforms. An alternative regulatory strategy when it comes to platforms involves the ex poste use of tort liability to incentivize "better" (by some metric) behavior. Platforms benefit from a form of intermediate immunity from tort liability under Section 230 of the Communications Decency Act.[135] Section 230 immunity is likely wider than the immunity from liability available under the First Amendment,[136] although the scope of

which will often be to silence their critics and to suppress ideas they do not like").

132. Elena Kagan, *Private Speech, Public Purpose: The Role of Governmental Motive in First Amendment Doctrine*, 63 U. CHI. L. REV. 413, 414 (1996).

133. Lum & Lazovich, *supra* note 26.

134. For a nuanced account of the difficulty of curbing the "bad tendencies of democracy," see David A. Strauss, *Corruption, Equality, and Campaign Finance Reform*, 94 COLUM. L. REV. 1369, 1378–79 (1994).

135. 47 U.S.C. § 230; *see also* Zeran v. Am. Online, Inc., 129 F.3d 327, 328 (4th Cir. 1997) (holding that Section 230 immunized an online service provider from liability for content appearing on its site created by another party).

136. *Cf.* Note, *Section 230 as First Amendment Rule*, 131 HARV. L. REV. 2027, 2030 (2018) (noting that "[j]udges and academics are nearly in consensus in assuming that the First Amendment does not

constitutionally permissible tort liability remains incompletely defined.[137]

Recent lawsuits have tried to pierce Section 230 immunity from various angles. Some have tried to exploit federal statutory liability for aiding and abetting political violence.[138] Others lean on common law tort theories, but contend that Section 230 does not extend to suits that turn on platforms' use of algorithmic controls to sequence and filter content. For example, in an August 2024 decision, a panel of the Third Circuit reversed a district court's dismissal of a common law tort complaint against TikTok for its promotion of content that played a role in the death of a minor.[139] The circuit court held that Section 230 did not extend to a claim that TikTok's "algorithm was defectively designed because it 'recommended' and 'promoted' the Blackout Challenge."[140] The Blackout Challenge, said the panel, was "TikTok's own expressive activity," and as such fell outside Section 230's scope.[141] This construction of Section 230 has been severely criticized.[142] Thus, it is far from clear how this ruling can be squared with the seemingly unambiguous Section 230 command that no platform can "be treated as the publisher or speaker of *any* information provided by another information content provider."[143]

Reflection on the prospect of tort liability that is delimited in this fashion and consistent with Section 230 (especially with the idea of "islands of algorithmic integrity" in mind) offers some further reasons for skepticism of the Third Circuit's decision and the consequences of tort liability for algorithmic design more generally. For it is far from clear how algorithmic-design-based liability of the sort that the Third Circuit embraced can be cabined. Every algorithmic decision changes the overall mix of content on the platform. So, it is always the case that such decisions in some sense "cause" the appearance of objectionable content.[144] Indeed, one could argue that any mechanism imposed to limit one sort of harmful speech necessarily increases the likelihood that other sorts of speech (including other sorts of

require § 230").

137. Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2046 (2018).

138. *See, e.g.*, Twitter, Inc. v. Taamneh, 598 U.S. 471, 503 (2023) (rejecting that reading of federal statutory tort liability).

139. Nylah Anderson watched a TikTok video on the "Blackout Challenge" and died imitating what she saw. Anderson v. TikTok, Inc., 116 F.4th 180, 181 (3rd Cir. 2024).

140. *Id.* at 184.

141. *Id.*

142. *See, e.g.*, Ryan Calo, *Courts Should Hold Social Media Accountable—But Not By Ignoring Federal Law*, HARV. L. REV. BLOG (Sept. 10, 2024), https://harvardlawreview.org/blog/2024/09/courts-should-hold-social-media-accountable-but-not-by-ignoring-federal-law [https://perma.cc/CFE6-3ZDZ].

143. 47 U.S.C. § 230(c)(1) (emphasis added).

144. One might interpose here some notion of algorithmic proximate cause. That presents, to say the least, rather difficult questions of doctrinal design.

harmful speech) will feature prominently on the platform. For example, a decision to filter out speech endorsing political violence is (all else being equal) going to increase the volume of speech that is likely conducive to adolescent mental health problems. In this way, the Third Circuit's decision (at least as written) has the practical effect of carving out *all* algorithmic content-moderation activity from Section 230's scope. It is hard to imagine this concurs with Congress's enacting intent.

Indeed, tort liability for algorithmic decision will inevitably push platforms to rely more on networks, rather than algorithms, as drivers of content. But the empirical evidence suggests that network-based platform designs are more, not less, likely to experience higher levels of fake news, and that they are less amenable to technical fixes.[145] Tort liability, at least as understood by the Third Circuit in the TikTok case, therefore pushes platforms *away* from socially desirable equilibria. Paradoxically, all else being equal, it is likely to increase, and not decrease, the volume of deeply troublesome material on platforms of the sort at issue in the Third Circuit TikTok case. More generally, it is again hard to see how liability for algorithmic design decisions, all else being equal, is socially desirable.

B. THE POSSIBLE VECTORS OF ALGORITHMIC INTEGRITY

The fact that state and national governments opt for partial or unwise regulatory strategies does not mean that are no promising paths forward. To the contrary, the examples examined in Part II suggest a range of useful reforms. I outline three here briefly.

To begin with, the examples of Wikipedia and the BBC suggest that it may be possible to build at least small-scale islands of algorithmic integrity either in the private or the public sector. Those examples further suggest that whether state or private in character, such an island needs mechanisms to shield itself from the pressure to maximize profits. An entity that is exposed to the market for corporate control is unlikely to be able to resist commercial pressures for long.

Corporate form hence matters. For example, social platforms' incentive to maximize engagement, and hence maximize advertising revenue, has been "critical" to driving the dissemination of radicalizing and hateful speech.[146] The transformation of Twitter to X after its purchase by Elon Musk, and the subsequent degradation and coarsening of discourse on the platform, offer an abject lesson in the perils of the unfettered free market for islands of

145. *See supra* text accompanying notes 44–65.
146. DARON ACEMOGLU & SIMON JOHNSON, POWER AND PROGRESS 362 (2023).

algorithmic integrity.[147] The market for corporate control, which is often glossed over in light of the efficient capital markets hypothesis, is commonly viewed as an unproblematic good.

One of the main lessons of the islands of integrity literature, however, is the need for well-motivated leadership of the sort that has been described at Wikipedia and the BBC. It is hard to see how such motivation survives under the shadow of potential corporate takeover.

Second, islands of integrity require the right means (or tools), as well as the right motive. The use of algorithmic tools to curate a platform creates means in a way that reliance on network effects does not. It is thus a mistake to assume, as the Third Circuit seems to have done in the TikTok case, that an algorithmically managed platform is worse than a network based one. As Part I illustrated, the empirical evidence suggests that algorithmically managed platforms are generally not more polluted by misinformation than ones driven by users' networks.[148] Quite the contrary.

Moreover, a social platform built around an algorithm may have tools to improve its epistemic environment that a network-based platform lacks. For instance, a 2023 study found that certain "algorithmic deamplification" interventions had the potential to "reduce[] engagement with misinformation by more than [fifty] percent."[149] Another example of an instrument for epistemic integrity is, somewhat surprisingly, a feature of Facebook's algorithm, which has baked in a preference for friends-and-family content that "appears to be an explicit attempt to fight the logic of engagement optimization."[150]

Third, there is a range of tailored reforms that precisely target ways in which social platforms stand in asymmetrical relations of exploitation and dominance to their users. As a very general first step, Luca Belli and Marlena Wisniak have proposed the use of "nutrition labels," detailing key parameters of platform operation as a way of enabling better informed consumer choice between platforms.[151] This kind of notice-based strategy,

147. There is some evidence that X systematically favored right-leaning posts in late 2024, suggesting a link between corporate control and political distortion. Timothy Graham & Mark Andrejevic, *A Computational Analysis of Potential Algorithmic Bias on Platform X During the 2024 US Election* (Queensland Univ. of Tech., Working Paper, 2024)), https://eprints.qut.edu.au/253211.

148. Budak et al., *supra* note 52, at 48; *accord* Hosseinmardi et al., *supra* note 47, at 1.

149. Benjamin Kaiser & Jonathan Mayer, *It's the Algorithm: A Large-Scale Comparative Field Study of Misinformation Interventions*, KNIGHT FIRST AMEND. INST. (Oct. 23, 2023), https://knightcolumbia.org/content/its-the-algorithm-a-large-scale-comparative-field-study-of-misinformation-interventions [https://perma.cc/Y4KU-76BY].

150. Narayanan, *supra* note 10, at 31.

151. Luca Belli & Marlena Wisniak, *What's in an Algorithm? Empowering Users Through Nutrition Labels for Social Media Recommender Systems*, KNIGHT FIRST AMEND. INST. (Aug. 22, 2023),

while plausible to implement, assumes a measure of user choice over which platform to use. At present, such choice is largely illusory because of the market dominance of a small number of platforms.[152] It is also hard to see how consumers, particularly those already at the ideological margin, could be persuaded to make the right kind of choice. Inducing more competition, and hence more consumer choices, in social platforms would give notice-oriented measures more bite. Some work has been done on potential varieties of platform design,[153] but there remains ample room for inquiry and improvement. The basic point, though, is that some combination of increased competition and better consumer-facing notices would better allow certain users to select among different social platforms based on their own preferences—although it is hard to be confident that the right users, so to speak, will be those aided.

There are also steps that can be taken by a well-motivated platform manager. Within a platform, for example, the BBC's strategy of promoting personalization could be adopted and redeployed in a number of ways. For instance, bots, or "user-taught" agents could be supplied to help individual users curate the shape of their feeds over time.[154] These bots, however, might be constrained by the understanding of the platform's mission, which excluded normatively troublesome activity characterizing the tails of the ideological distribution.

Finally, another way of mitigating exploitation concerns focuses on advertisers rather than users. Firms advertising on platforms are often unaware their products or services are marketed next to fake news, despite having an aversion to that arrangement.[155] They lack, however, information on when and how this occurs. Increased disclosure by platforms on "whether . . . advertisements appear on misinformation outlets," as well as increased "transparency for consumers about which companies advertise" there, provides the potential to stimulate a collective shift to a more truthful

https://knightcolumbia.org/content/whats-in-an-algorithm-empowering-users-through-nutrition-labels-for-social-media-recommender-systems [https://perma.cc/N7MW-SEVT].

152. Lina M. Khan, *The Separation of Platforms and Commerce*, 119 COLUM. L. REV. 973, 976 (2019) ("A handful of digital platforms exert increasing control over key arteries of American commerce and communications.").

153. For a recent survey of other possible models of "decentraliz[ed]" platform governance, see Ethan Zuckerman & Chand Rajendra-Nicolucci, *From Community Governance to Customer Service and Back Again: Re-Examining Pre-Web Models of Online Governance to Address Platforms' Crisis of Legitimacy*, 9 SOC. MEDIA + SOC'Y, July–Sept. 2023, at 1, 7–9.

154. Kevin Feng, David McDonald & Amy Zhang, *Teachable Agents for End-User Empowerment in Personalized Feed Curation*, KNIGHT FIRST AMEND. INST. (Oct. 10, 2023), https://knightcolumbia.org/content/teachable-agents-for-end-user-empowerment-in-personalized-feed-curation [https://perma.cc/RAN8-QT7S].

155. Wajeeha Ahmad, Ananya Sen, Charles Eesley & Erik Brynjolfsson, *Companies Inadvertently Fund Online Misinformation Despite Consumer Backlash*, 630 NATURE 123, 125–28 (2024).

equilibrium.[156] Such disclosures help ensure that "the means of ensuring legibility [will not completely] fade into the background of the ordinary patterns of our li[ves],"[157] as platform affordances become too banal to notice. Such disclosures, finally, might be mandated by law, potentially as a means of mitigating fraud concerns related to platform use.

CONCLUSION

In this essay, I have tried to offer an affirmative vision of social platform governance in the long run, or at least the seeds of such a vision. No doubt this vision is leagues away from the grubby, venal, and hateful reality of social platforms now. It is, indeed, a stark contrast to those extant realities. But one of the functions of scholarship is to generate plausible pathways away from a suboptimal institutional status quo. The articulation of alternatives is itself of value.

As I have suggested, drawing on sociological and political science literature on islands of integrity in public administration allows us to see some of the limits of existing regulatory strategies with respect to social platforms. Doing so opens up new opportunities for improved public and private governance. Of course, the model of islands of integrity in a public administrative context cannot be mechanically transposed over to the platform context. But by offering us a new North Star for reforming governance efforts, I hope it can advance our understanding of how to build platforms fit for our complex, yet (perhaps still) fragile democratic moment.

156. *Id.* at 129.

157. Henry Farrell & Marion Fourcade, *The Moral Economy of High-Tech Modernism*, 152 DÆDALUS 225, 228 (2023).