SYSTEMIC SOCIAL MEDIA REGULATION

FRANK FAGAN[†]

ABSTRACT

Social media platforms are motivated by profit, corporate image, long-term viability, good citizenship, and a desire for friendly legal environments. These managerial interests stand in contrast to the gubernatorial interests of the state, which include the promotion of free speech, the development of e-commerce, various counter terrorism initiatives, and the discouragement of hate speech. Inasmuch as managerial and gubernatorial interests overlap, a self-regulation model of platform governance should prevail. Inasmuch as they diverge, regulation is desirable when its benefits exceed its costs. An assessment of the benefits and costs of social media regulation should account for how social facts, norms, and falsehoods proliferate. This Article sketches a basic economic model. What emerges from the analysis is that the quality of discourse cannot be controlled through suppression of content, or even disclosure of source. A better approach is to modify, in a manner conducive to discursive excellence, the structure of the forum. Optimal platform architecture should aim to reduce the systemic externalities generated by the social interactions that they enable, including the social costs of unlawful interference in elections and the proliferation of hate speech. Simultaneously, a systemic approach to social media regulation implies fewer controls on user behavior and content creation, and attendant First Amendment complications. Several examples are explored, including algorithmic newsfeeds, online advertising, and invited campus speakers.

I. INTRODUCTION

Recently, during the ongoing wave of sexual assault claims involving celebrities, newscasters, politicians, and high-profile Hollywood employers, NPR's *All Things Considered* ran a segment entitled, "Women Are Speaking Up About Abuse, But Why Now?"[1] Why now is a good question. Sexual harassment has only been recognized as an

---

[†] Associate Professor of Law, EDHEC Business School, France. The author would like to thank Saul Levmore for comments.

[1] *All Things Considered: Women Are Speaking Up About Abuse, but Why Now?* NAT'L. PUB. RADIO (Oct. 27, 2017), https://www.npr.org/2017/10/27/560231232/women-are-speaking-up-about-harassment-and-abuse-but-why-now.

actionable form of discrimination just recently; socio-cultural advances in gender equality are relatively new phenomena.[2] If anything, prior levels of abuse were at least as high as they are today—if not higher. So what explains the recent spike in claims? During the segment, Anita Hill conjectured that, "[since passage of the Civil Rights Act of 1991], we have been raising children—daughters in particular—with the understanding that sexual harassment is illegal, shouldn't be tolerated, and that it's wrong."[3] In other words, law drives morality and norms, and it has finally led to concrete action twenty-five years later. Professor Hill is right to point out that lawmakers can shape social norms, which can lead to waves of litigation a generation delayed.[4] Greater numbers of women are in college, working, and earning higher wages, and all of this too, has played a role. But is social media particularly adept at creating tipping points, and if so, how exactly does social media accelerate change?

Legal scholars have explained how expressive laws, such as the Civil Rights Act of 1991, check bad behaviors that are difficult or costly to detect.[5] New laws that express social values make it easier for victims to speak out against social wrongs, lowering overall enforcement costs; they raise public awareness of social problems, increasing the number of people who sanction a wrongdoer; and shine spotlights on particular issues, increasing the probability that public enforcement resources will be allocated in the first place. And at times, new laws actually define what is morally right and wrong *a priori*. Lawmakers function as "norm entrepreneurs" who proclaim and calibrate social values, intuitions, and morality.[6] Normative claims, embodied in new laws, can generate "norm cascades" and "norm bandwagons," which quickly lead to new forms of social behavior.[7] Cass Sunstein, Richard McAdams, Bob Cooter, and others have written extensively on how these patterns unfold and what they mean for law.[8]

---

[2] *See* CATHERINE MACKINNON, WOMEN'S LIVES, MEN'S LAWS 34–43 (2007) (documenting the recognition of sexual harassment and assault as legal claims of sexual discrimination).

[3] NAT'L. PUB. RADIO, *supra* note 1.

[4] *See infra* § IV.B.

[5] A good example is Lawrence Lessig, *The Regulation of Social Meaning*, 62 U. CHI. L. REV. 943 (1995).

[6] *See id.* at 1019 (providing examples).

[7] *See*, *e.g.*, Cass R. Sunstein, *Social Norms and Social Roles*, 96 COLUM. L. REV. 903, 912 (1996) (explaining that rapid shifts in new norms help explain the attack on apartheid in South Africa, the fall of Communism, the election of Ronald Reagan, and the rise of the feminist movement).

[8] *See*, *e.g.*, *id.*; Cass R. Sunstein, *Unleashed*, Aug. 22, 2017,

This Article aims to add to that body of work in several ways. First, it incorporates the role of social media into the analysis of norm evolution and explains its bearing on law. While the analysis of analog social networks is as old as society itself,[9] their digital counterparts generally accelerate, and in some cases make possible, the cascade and bandwagon effects described by Sunstein and others. To support this claim, this Article develops a basic economic model that describes the creation and proliferation of social facts, norms, and falsehoods, general enough to account for citizens, lawmakers, and social media platforms acting as norm entrepreneurs. By working through the details, this Article provides a clear rationale for systemic social media platform regulation. Several important policies are taken up, including how governments should regulate algorithmic news feeds and online advertising, and how universities should regulate the promotion of controversial speakers organized by students.

In brief, factual and normative evolution begins with a person or lawmaker who claims that some fact is true, or that a given behavior is morally wrong. Because social media platforms control the flow of content manufactured by people and lawmakers, platforms function as norm entrepreneurs as well, which, at least from a functional standpoint, is at odds with § 230 of the Communications Decency Act.[10] A key component of this Article, therefore, is to explore the communicative role of platforms by clarifying how claim proliferation depends on their architecture. What emerges from the analysis is that regulation, when warranted, should focus on platform architecture and not platform speech. This conclusion is consistent with § 230 inasmuch as platform liability would be based upon violations of systemic rules rather than responsibility for speech.

Regardless of who is responsible for the origination of a factual or normative claim, claims can spread leading to new patterns of behavior and social enforcement. For instance, lawmakers recognized sexual harassment as a social wrong when they passed the Civil Rights Act of 1991. Validity claims are also made by friends, law professors, media executives, foreign agents purchasing a Facebook ad—essentially anyone

---

https://papers.ssrn.com/abstract=3025749; RICHARD A. MCADAMS, THE EXPRESSIVE POWERS OF LAW: THEORIES AND LIMITS 42 (2015) (describing the informative and focalizing effects of norms and the relevance to law of these effects).

[9] Aristotle's *Politics* suffices for documenting the analysis and recognized importance of social networks in the ancient world. *See generally* ARISTOTLE, POLITICS (R.F. Stalley ed. Ernest Barker trans. Oxford Univ. Press Release ed. 2009) (c. 350 B.C.E.).

[10] *See infra* § IV.C.

who can be heard. When claims stick, they are replicated over social networks and unleash norm cascades.

The principal difference between analog and digital social networks is that the latter consist of broad and horizontal membership across narrow attitudes and beliefs. While the bonds between social media platform members are often thin, platform architecture supports greater volume and breadth when compared to traditional social networks. Norm entrepreneurs who effectively tap platform networks and make their claims "go viral," exercise marginal de facto rulemaking, fact-making, and potentially outsized influence in socio-legal activities such as national elections.[11] It is true that rumors, and remedies for injurious rumors, are as old as the wellsprings of common law,[12] but platforms can accelerate, and in some instances make possible, the proliferation of defamatory and other types of facts. Moreover, facts have lifecycles. They proliferate when the social network that they harness is systemically strong and resilient to counter-facts, falsehoods, and other forms of declaratory corruption.

The proliferation of a fact or norm, or a particular version of either, is valuable to the extent that it changes behavior, and this value can accrue to both private entities and the state alike. Social media platforms such as Facebook, YouTube, and Twitter pursue managerial interests, which include profit, corporate image, long-term viability, good citizenship, and friendly regulatory environments. These stand in contrast to gubernatorial interests, which are pursued by governments, and include the promotion of free speech, the development of e-commerce, counter terrorism initiatives, and the discouragement of hate speech.[13] Inasmuch

---

[11] In some cases, non-digital social networks can be large, narrow, and dense. Consider that Italy's top three television broadcasters, controlled by Silvio Berlusconi, were able to reach nearly the entire electorate with a narrow message that was consistently replicated. *See* Rachel Sanderson, *Berlusconi Study Sheds Light on Politics and Profits*, FIN. TIMES, (January 4, 2017), https://www.ft.com/content/eb86eb6c-d284-11e6-9341-7393bb2e1b51 (documenting Berlusconi's control of Mediaset and its use to continually message a national audience during his campaigns for Prime Minister). The point is that social media platforms generally exhibit these structural features.

[12] *See* J.H. BAKER, AN INTRODUCTION TO ENGLISH LEGAL HISTORY 436-46 (4th ed. 2002) (noting that defamation torts were subject to the jurisdiction of English ecclesiastical courts prior to the 17th century and then transferred to common law courts).

[13] To be sure, platforms engage in private governance when they choose to suppress or permit speech. *See* Jack Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 UC DAVIS L. REV. 1149, 1153 (2018) (noting that a practical ability to engage in

as managerial interests and gubernatorial interests overlap, self-regulation maximizes social welfare so long as the interests themselves are efficient.[14] Inasmuch as they diverge, regulation can be socially useful if its benefits exceed its costs. Some of the features of current proposals, such as the one advanced by Senators Klobuchar, Warner, and McCain[15] (KWM) intuitively track this approach. One of the aims of this Article is to provide some theoretical underpinnings for the KWM proposal, and similar ones, grounded in microeconomics and social psychology.

Assertions of factual or moral validity, when systemically spread throughout social media, can eventually lead to unconscious compliance and instinctual self-regulation. While this type of evolutionary path takes a substantial period of time to complete, construction of deep policies and normative legacies may be the only path available to lawmakers in polarized political environments; or when enforcement costs of legal rules are so excessive, lawmakers may have no options for generating compliance other than nurturing moral instinct.[16] These longer-term pathways toward deep policy and unconscious compliance can help explain attitudes toward immigration, climate change, and religious expression in contexts such as the creation of wedding cakes for celebrating same-sex marriages.[17] It should be obvious that the construction of deep policy is not the only way to shape social attitudes or

---

speech acts is subject to the private decisions of platform owners, which amounts to private governance). For clarity, this Article treats platform governance, or private content moderation, as a means by which platforms pursue their managerial interests. For instance, a platform may suppress hate speech in order to attract and keep users or avoid regulatory interference. When doing so, this Article considers that platform as pursuing its managerial interests.

[14] Recent scholarship suggests that policymakers should recognize the intricate content moderation systems already in place and concludes that law should take self-regulation efforts into account when considering policy. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1666 (2018) ("[A]ny proposed regulation . . . . should work with an understanding of the intricate self-regulatory structure already in place . . ."); Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEORGETOWN L. J. *11 (2018) ("Whether we decide to regulate platform censorship or leave it to the market, the decision should be considered").

[15] The Honest Ads Act, § 1989, 115th Cong. (2017).

[16] *Cf.* FRANK FAGAN, LAW AND THE LIMITS OF GOVERNMENT: TEMPORARY VERSUS PERMANENT LEGISLATION 105 (2013) (providing empirical support for the notion that the difficulty of passing laws in some domains is more difficult than others).

[17] *See infra* § III.B.3.

obtain desired legal outcomes. In some instances, however, construction of deep policy is possible and defensible within a welfarist, rational-choice framework. Exploring and developing this approach has important implications for the regulation of social media. To the extent that long-term managerial interests consist of the development, moderation, and shaping of public opinion over time, the government has a weaker basis for platform regulation, at least if that basis is grounded in First Amendment doctrine.[18] By contrast, when the benefits of speech suppression are certain and immediate, the basis for platform regulation is stronger, though direct confrontation with First Amendment principles can be avoided through a systemic regulatory approach.

Social norms, of course, have historically been the domain of sociologists, who have described compliance behavior in contrast to consequentialist conceptions of rules as prices.[19] For them, rules can operate in gray areas as integral parts of social processes that prompt self-regulation either through stabilizing existing patterns of rule-based behavior, or through transforming widespread ideas of what is the appropriate mode of interaction.[20] Surprisingly, a number of contemporary economists view compliance as a result of some mix between regulatory deterrence on the one hand, and relational obligation, socialization, reputation, legitimacy, and other phenomena stemming from social interaction on the other.[21] Even within law and economics, whose consequentialist tradition naturally embraces "expected punishment" explanations of compliance, there has arisen a large project aimed at understanding the interaction between social norms and the law.[22]

---

[18] *Infra* § IV.A.

[19] *See*, *e.g.*, EMILE DURKHEIM, THE RULES OF SOCIOLOGICAL METHOD 2 (8th ed. 1938) (contrasting obedience with the law to the coercive power of moral maxims and "the public conscience [which] exercises a check on every act which offends it by means of the surveillance it exercises over the conduct of citizens").

[20] *See*, *e.g.*, MICHEL FOUCAULT, THE HISTORY OF SEXUALITY, VOL. 2 144 (1990) (theorizing that modern compliance regimes achieve success by transforming ideas of appropriateness).

[21] *See* Jon G. Sutinen & K. Kuperan, *A Socio-Economic Theory of Regulatory Compliance*, 26 INT'L J. SOCIO-ECON. 174, 175 (1999) (presenting a model of compliance behavior where rational agents are motivated extrinsically and intrinsically); GARY S. BECKER, ACCOUNTING FOR TASTES 162 (1998) (developing a model of economic behavior where agents derive utility from social interactions).

[22] See the essays collected in SOCIAL NORMS, NON-LEGAL SANCTIONS, AND THE LAW (Eric A. Posner ed. 2007). See also the *Journal of Legal Studies*

Nonetheless, leading theory in law and economics can sometimes lack clarity and full elaboration. It often provides a muddled price-theoretic explanation for adherence to social norms: people derive consequential utility from acting virtuously, even when their virtuous actions are so obviously averse to their interests along other dimensions.[23] For example, a generosity norm confers utility on practitioners of self-sacrifice and altruism, or a norm of environmental stewardship confers "psychic" utility even though stewardship may be monetarily costly and offer few individual rewards in both the short and long run. Nearly always, foundational norms like generosity, and others that generate psychic utility, are described as preexisting or rooted in evolutionary biology.[24] That law and economics scholars have had little to say about how foundational norms can be created by instrumentally driven agents may be due to an intuition that guiding unconscious social norms toward normative ends with deep policy can be easily short-circuited with conscious rules; or perhaps it is due to the idea that welfarist guidance of instinctual behavior may otherwise have little payoff in terms of social value because of time and discount rates.[25]

---

symposium, *Social Norms, Social Meaning, and the Economics Analysis of Law*, 27 J. LEG. STUD. 537–823. An important early contribution that describes the ability of law itself to create and modify social norms is Robert Cooter, *Expressive Law and Economics*, 27 J. LEGAL STUD. 585 (1998). *See also* Robert Cooter, *Do Good Laws Make Good Citizens?*, 86 VA. L. REV. 1577 (2000). For a recent restatement and expansion of the theory and its limits, see generally MCADAMS, *supra* note 8.

[23] This approach has invited numerous critiques and outright dismissal. *See*, *e.g.*, Elizabeth Anderson, *Beyond* Homo Economicus*: New Developments in Theories of Social Norms*, 29 PHIL. & PUB. AFF. 170, 193 (2000) (arguing that a desire for identification with a group can explain the motive to comply with a norm).

[24] One strategic explanation for the existence of other-regarding behaviors that run contrary to self-interest is that they function as impulse-control devices and signal commitment to shared long-term goals with potential partners. Thus, moral emotions and feelings are instrumental towards achieving particular ends. *See* ROBERT H. FRANK, PASSIONS WITHIN REASON: THE STRATEGIC ROLE OF THE EMOTIONS 211 (1988). On how evolutionary biology has shaped norms, see PAUL BLOOM, AGAINST EMPATHY 170 (2017) (noting that "[w]e are naturally kind because our ancestors who were kind to others outlived and outreproduced those who [were not]," and that "[this] doesn't mean that when people help others they are thinking about survival and reproduction").

[25] Note that this thinking closely tracks First Amendment doctrine and its apathy toward suppressing speech on the basis of non-immediate effects. *See infra* § IV.A.

Even if either of those intuitions are true, an instrumental explanation for the origin of unconscious normative behavior can bring coherence to consequentialist theories of social norms and can present opportunities for deep social welfare maximization. Rather than stripping social norms (and their attendant morality) of their independence and reducing them to a master concept of utility, providing an instrumental account actually has the opposite effect here. Deeper understanding of the pathways toward unconscious behavior, cast in rational-choice terms, frees up conceptual space for the construction of deep policies that elicit unconscious compliance while respecting self-interest precisely because the interests are constructed with citizen input.[26] Moreover, a deeper understanding of the micro-foundations of social contagion is fundamental for addressing the systemic risk generated by social media.

The key is to develop a thorough understanding of how platforms shape social facts and norms. Platforms obviously represent important pathways into conscious opinion formation, but existing proposals for intervention (or self-regulation) can benefit from explicitly acknowledging the deeper, unconscious attitudes and behaviors in play. By unpacking how platforms systemically generate attitudes and beliefs through social channels, linkages, sequesters, and quarantines, the externalities generated by social media come into full focus.

Providing a detailed theory of the creation of social facts, norms, and falsehoods may give the impression of excess or intellectual overindulgence. On the other hand, a theory of social behavior, with respect to media, obviously seems necessary for developing a systemic approach to the regulation of social media—especially one that avoids being under-inclusive, but perhaps more importantly, avoids being over-inclusive by taking seriously the shaky claim that impostors and dissemblers can control public opinion and social facts.[27] Norm entrepreneurs require more than just a low-cost platform that can reach many people. They must make claims that resonate throughout a social network and trigger sticky patterns of approval, disapproval, pride, and

---

[26] Interesting complications for welfare analysis arise when the democratic input of a current generation binds a future one. *See* Frank Fagan & Saul Levmore, *Legislative Sunrises: Transitions, Veiled Commitments, and Carbon Taxes*, FRANK FAGAN & SAUL LEVMORE, THE TIMING OF LAWMAKING 130, 133–37 (2017).

[27] *Cf.* FRANKLIN FOER, WORLD WITHOUT MIND: THE EXISTENTIAL THREAT OF BIG TECH (2018) (noting that "if we want to be melodramatic about it, we could say Facebook is constantly tinkering with the quality of news and opinion that it allows to break through the din, adjusting the quality of political and cultural discourse").

guilt that burrow deeply into the social experience.

II. SOCIAL NORMS AND SOCIAL MEDIA

A. Attitudes and Compliance in a Vacuum

Compliance behavior is created. Typically, in law, creation is more or less immediate. Lawmakers produce rules backed by sanctions, which generate rule-following behavior.[28] Otherwise, creation takes time, and in some cases, so much time, that a proclivity toward compliance appears endowed or the result of a protracted evolutionary path.[29] While recent advances in theoretical biology acknowledge that we "think fast", and that fast thinking occurs independently of slow rational-choice-style thinking,[30] as will become clear, none of this has much to do with deeply ingrained and morally charged norms like generosity toward migrants and strangers, stewardship toward the environment, or social attitudes toward the separation of church and state. People do not think fast about these issues. For a given set of relations among people, and a given span of time, the underlying claims and rationales for thinking a particular way either operate primarily in the socially cognitive foreground, where discourse and messaging explicitly recognize the role that norm entrepreneurs and social networks play in shaping group attitudes and beliefs, or they operate in socially cognitive backgrounds where inconspicuous claims and rationales can, oftentimes, continue to generate strong compliance and attitude-shaping effects. Put differently, compliance and beliefs can be conscious or unconscious, but in either case, they are the fruits of utility-seeking norm entrepreneurs who have leveraged social networks.[31]

---

[28] For this discussion, it is useful to set aside other purposes for rules, such as enabling cooperation and coordination through bodies of law like contracts, corporations, trusts, and estates and to focus exclusively on generating compliance with commands backed by threats. The distinction between law as a command and law as a coordination device is elaborated in H.L.A. HART, THE CONCEPT OF LAW 26 (2d ed. 1961).

[29] For instance, Adam Smith notes that:

> [M]en are naturally endowed with a desire for the welfare and preservation of society; but the Author of nature hasn't left it to men to use their reason to work out what kinds and levels of punishment are right for this purpose; rather, he has endowed men with an immediate and instinctive approval of just precisely the kind and level of punishment that is most proper to attain it.

ADAM SMITH, A THEORY OF MORAL SENTIMENTS II.i.5.10 (1759).

[30] DANIEL KAHNEMAN, THINKING, FAST AND SLOW 19 (2011).

[31] Throughout this Article, social networks refer to online and offline groups of people who are connected through social contact and exhibit some level of

A social attitude toward immigrants, monuments, single-payer healthcare, wealth disparity, and other politically charged issues begins, at some point in time, with a single claim made by a norm entrepreneur.[32] This is true for attitudes toward contestable facts as well. The norm entrepreneur might be a law professor, an acquaintance or parent, a lawmaker, an evangelist, an industry leader, or an agent for a foreign power.[33] Norm entrepreneurs make validity claims through a medium. For instance, the law professor may claim that immigration policy should embrace open borders through a lecture, or a parent may claim that generosity should be shown only toward friends and family during dinner-table conversation. Likewise, foreign agents may claim that a particular candidate is corrupt through a Facebook ad. Inasmuch as social media platforms control the flow of validity claims, they too, function as norm entrepreneurs.

In each instance, the norm entrepreneur makes a claim that attempts to persuade the listener of its validity. In a vacuum, the strength or magnitude of validity is based exclusively on the listener's existing beliefs.[34] If the student holds weak personal views on immigration and

---

density greater than zero. For a given context, social media refers to online social networks (groups of people), the content contained on those networks, and the platforms which provide the space for the people to connect and share content.

[32] *See* JÜRGEN HABERMAS, A THEORY OF COMMUNICATIVE ACTION, VOL. 1 8–9 (1981) (theorizing that norm entrepreneurs make "validity claims," or assertions that are accepted or not accepted as true by other individual persons and that claims are ultimately validated or invalidated by a society through a series of communicative actions). For an early elaboration of this idea, see SMITH, *supra* note 29 at I.i.3.2 (noting that to approve or disapprove of another's opinion is to adopt or not adopt those opinions).

[33] For the proposition that lawmakers create social norms, see Lessig, *supra* note 5 at 1019; *see also* Emmanuela Carbonara, et al., *Lawmakers as Norm Entrepreneurs*, 4 REV. L. & ECON. 779, 779 (2008) (developing an economic model of legislative norm entrepreneurship); *infra* § IB. V. For the proposition that social media platforms create social norms, see *infra* § IV.C.

[34] Thus, a person's susceptibility to a claim, at a given moment in time, is based upon all of the previous claims that she has processed and how those claims have been refined and shaped by her experiences. Younger people are generally more impressionable because they have processed less claims. *See* SMITH, *supra* note 29 at I.i.3.10 (noting that people evaluate the communicative acts of others through comparisons to their own experiences and thinking). Note that this approach is consistent with contemporary theories of the mind. Materialist theories that understand all human experience as the result of electrical activity in the brain, and view all output and thinking as the result of stimulus and input

believes that the law professor is authoritative and trustworthy, then she may accept the claim as true. The same claim made by a long-standing acquaintance may have less (or more) persuasive power. On the other hand, the authority of the norm entrepreneur may matter little. If the Surgeon General makes a claim that smoking leads to bad health, but supports that claim with state-of-the-art scientific evidence, then the listener may accept that claim as valid despite what she thinks of the Surgeon General. Similarly, if an accomplished climatologist makes a claim that global surface temperatures are rising, a voter may reject that claim despite what she thinks of the climatologist or the scientific evidence.[35] Much like a judge considering the validity of a fact or expert testimony, a person considering the validity of a claim made by a norm entrepreneur evaluates its truth based upon its substance and the credibility of the speaker, and that evaluation itself is set against a backdrop of existing beliefs of what is true substance and what constitutes a reliable signal of credibility.[36]

Once a claim is made, a listener may reject it outright, and the claim may die immediately. The recipient, for example, may simply reject

---

(just like a computer), are perfectly compatible with the idea that people process claims based upon the sum of their previous experiences. *See generally*, STANISLAS DEHAENE, CONSCIOUSNESS AND THE BRAIN: DECIPHERING HOW THE BRAIN CODES OUR THOUGHTS (2014); STEPHEN PINKER, HOW THE MIND WORKS (1997). This approach also leaves space for theories that presuppose a preexisting genetic memory or naturally endowed morality; people would simply process validity claims against the backdrop of their current moral-mental state. *See* PAUL BLOOM, JUST BABIES: THE ORIGINS OF GOOD AND EVIL 218 (2015) (noting that we are born with "empathy and compassion, the capacity to judge the actions of others, and even some rudimentary understanding of justice and fairness").

[35] *Cf.* Cary Funk, *How Much Does Science Knowledge Influence People's Views of Climate Change and Energy Issues?*, PEW RESEARCH CTR. (March 22, 2017), http://www.pewresearch.org/fact-tank/2017/03/22/how-much-does-science-knowledge-influence-peoples-views-on-climate-change-and-energy-issues/ (noting that heightened science knowledge may have no discernible effect on beliefs that climate change is due to human activity). Receptivity toward scientific claims are discussed *infra* at § III.B. 2.

[36] Automated content moderation works the same way. An algorithm makes programmed determinations of the appropriateness of user-posted content such as photos based upon an existing database of illegal images. *See* Klonick, *supra* note 14, at 1636–37. Human moderators follow internal rules for making determinations of the appropriateness of content that includes ambiguities. *Id.* at 1638–39. For instance, a photograph may be classified as nudity or art based upon existing beliefs of what constitutes art.

the claim that a presidential candidate is in bad health because it does not comport with her existing beliefs, and an anonymous norm entrepreneur making a claim through a dubious Facebook ad does not persuade her otherwise. Or she may hold no existing beliefs about the candidate at all, but does not consider any form of political advertisement credible, so that advertisements in general make no impression upon her. Possible explanations may be that she currently accepts that most political news is fake, foreign agents have interfered with social media, or people should ignore political messaging entirely. All of this leads her to remain immune to the claim of bad health made by the norm entrepreneur.

B. Attitudes and Compliance in a Network

While existing beliefs may cause a person to reject or accept claims in a vacuum, where disapproval, approval, guilt, and pride play no role, the calculus changes entirely once a claim passes through a person's social network. Once placed on a network, a validity claim generates second-party enforcement through social approval or disapproval, which in turn, generates conscious, voluntary compliance with rules or provokes the adoption of social attitudes.[37] Validity claims, and patterns of approval and disapproval, can additionally generate sources of guilt and pride. For now, it is useful to set guilt and pride aside given that second-party social sanctioning can generate compliance and shape social attitudes entirely on its own. For instance, a civic leader may claim that strict border control is wrong because the United States is an immigrant nation, a model of humanitarianism, that it can lead more strongly through benevolence, and that the socio-economic benefits from enhanced diversity inure to society broadly. People who speak out, against immigration, can experience social approval or disapproval insofar as members of their social network approve or disapprove of the civic leader's position. Approval and disapproval depend upon the systemic composition of one's social network. Thus, demonstrated opposition to a claim can carry benefits from peer-approval and costs from peer-disapproval, and these are determined by the speaker's social network.[38]

---

[37] Thus, internationalization of claims and social sanctioning is not necessary for generating compliance and the adoption of social attitudes. Insofar as internal and external sanctioning are aligned, the former amplifies the effects of the latter. This point is developed in Part III, *infra*.

[38] People attempt to influence the composition of their networks in order to maximize net approval benefits. *See* Sinan Aral, et al., *Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks*, 106 PROC. NAT'L ACAD. SCI. 21544, 21544 (2009) (finding evidence that similarities among people drives more than fifty percent of behavioral contagion in online

Faced with peer benefits and costs, a person's utility is plainly shaped by their choice to oppose or support a claim initiated by the norm entrepreneur.[39] Consider that the entrepreneur, perhaps again citing to an authoritative study, may claim that organized football is dangerous for brain health. Under the gaze of friends or outspoken strangers, one may experience little to no disapproval, and may actually be rewarded with approval for ignoring the claim and its supporting research.[40] Different configurations of social networks may generate compliance or non-compliance. While the son of a neurologist may have difficulty mustering the courage to join the high school team, the son of a Heisman Trophy winner may experience disapproval for willfully avoiding participation. Presumably the son cares about his father's approval, but his network is larger. He must contend with his mother, his classmates, and indeed everyone with whom he has contact that may approve or disapprove of his decision. Through feedback, his entire social network bears on his choice to participate to the extent that he values their net approval. If we divide his utility into two parts, the first a measure of his enjoyment from playing football independent of his relationships with others, and the second a measure of enjoyment from playing strictly derived from his personal relationships, it is easy to see how his social network can change the outcome of his choice.

1. Speaking Out: Sexual Harassment in a Network

Similarly, the recent spike in public allegations of sexual harassment and abuse brought against public figures can be explained, in part, through the rise of social media platforms. While threats of retaliation, secret settlements, and other forms of discouragement can suppress claims, social media encourages them. Socially isolated claims, as opposed to socially networked claims, carry less probability of success. In isolation, no discernible pattern of misconduct can emerge. In a vacuum, perpetrators face no external consequences from others or law, which increases their capacities for retaliation, and in turn, chills claims. Even within a social network, retaliation can occur to the extent that the incident itself is quarantined and its validation relies on the credibility of the victim. Powerful perpetrators, acting within analog social networks where approval and disapproval flow more vertically and less horizontally

---

networks).

[39] Note that a norm entrepreneur can change the calculus of actual behavior, which can lead to real compliance with a rule and not just vocal support of a contested policy.

[40] Lawrence Lessig discusses a similar example with hockey helmets in Lessig, *supra* note 5 at 967.

amongst its members, can marshal social interactions to their advantage.[41] Network members who value the approval of perpetrators will refuse to validate the claim of a victim and may even make efforts to protect or support the perpetrators.[42] When perpetrators cannot resort to vertical social networks, they can rely on secret settlements that prohibit the victim from publicizing a claim. Commentators point to the superior bargaining power of perpetrators and the likelihood that secret agreements consist of lop-sided terms.[43] An additional problem is that secrecy obscures patterns of misconduct that would otherwise be uncovered had a claim been fully litigated.[44] Because patterns remain hidden, other claimants are unable to assess the full value of secrecy and properly deter serial harassers.

Given continued fears of retaliation, the endurance of vertical social networks that reward enablers, and the use of confidential settlements to silence victims, consider again the question of "why the wave of claims now?" Sexual harassment has been condemned, especially in the work place, for at least several decades.[45] While lawmakers and parents have clearly functioned as norm entrepreneurs, the key to understanding the cascade of claims is the emergence of social media platforms which have flattened social network membership. Unlike vertical networks, which enable retaliation and suppression of claims,

---

[41] *See* Ronan Farrow, *From Aggressive Overtures to Sexual Assault: Harvey Weinstein's Accusers Tell Their Stories*, THE NEW YORKER, (Oct. 23, 2017) https://www.newyorker.com/news/news-desk/from-aggressive-overtures-to-sexual-assault-harvey-weinsteins-accusers-tell-their-stories (noting that Harvey Weinstein orchestrated negative press coverage of his accusers in order to impugn their credibility).

[42] *See, e.g.*, *id.* (noting how Weinstein would orchestrate meetings with multiple assistants known as "honeypots" and then dismiss them from the meeting in order to isolate a victim).

[43] *See* Daniel Hemel, *How Nondisclosure Agreements Protect Sexual Predators*, VOX (Oct. 13, 2017), https://www.vox.com/the-big-idea/2017/10/9/16447118/confidentiality-agreement-weinstein-sexual-harassment-nda.

[44] *See* Saul Levmore & Frank Fagan, *Semi-Confidential Settlements in Civil, Criminal, and Sexual Assault Cases*, 103 CORNELL L. REV. 311, 311 (2018) (noting this problem and suggesting that sunshine-in-litigation laws should mandate semi-confidentiality, or revelation of the facts of a settlement, but not its terms).

[45] *See*, *e.g.*, Civil Rights Act of 1991, Pub. L. No. 102-166 (providing rights to sue and collect compensatory and punitive damages for workplace sexual harassment); Jenson v. Eveleth Taconite Co. 139 F.R.D. 657, 667 (D. Minn. 1991) (certifying a class of alleged sexual assault victims); Oncale v. Sundower Offshore Services, 523 U.S. 75 (1998) (permitting claims against perpetrators of the same sex).

horizontal networks diffuse power by multiplying sources of social sanctioning. These sources, while distributed broadly, provide immediate feedback that is disconnected from the influence of the perpetrator. Consider the victim's choice to publicize an episode of sexual harassment. She may fear retaliation, but the attendant emotional and financial costs are more likely to be balanced by social support and the validation of the truth of her claim. Social media acquaintances reiterate her experience through redundant messaging, sometimes anonymously, which in each instance, confers approval benefits and validates her claim. Moreover, existence of a horizontal network increases the likelihood that she can find meaningful work beyond the reach of the perpetrator's influence and can avoid being blacklisted. Note, too, that the social network which provides these benefits to her can weaken the vertical influence of the perpetrator. Enabling assistants and dishonest publicists face greater levels of disapproval should their support for the harasser become known.[46]

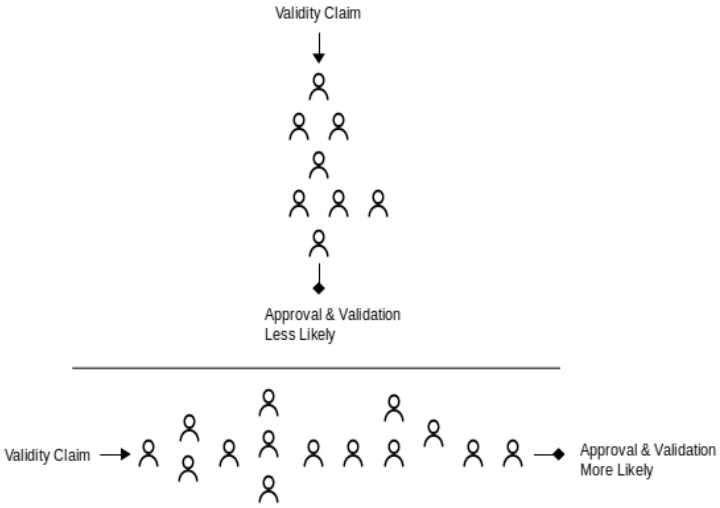

Figure 1: Passage of Validity Claims through Vertical and Horizontal Networks

Several decades ago, most victims could only make validity claims over analog networks, which more often exhibit vertical membership and are more susceptible to control by the injurer as a result.

[46] This is true to the extent that they are members of the horizontal network that is willing to sanction them. Broad network membership is a necessary condition for broad social sanctioning.

With the emergence of social media platforms, victims can pass validity claims through horizontal networks, which have increased the magnitude of approval benefits for speaking out and the likelihood that claims are validated. Reinforcing patterns of approval, enabled by horizontal network architectures, make it more likely for claims, and the norms that they embody, to go viral. The normative conclusion is that systemic rules that support horizontal network architectures are desirable when enforcement is costly, and in particular, when detection costs are excessive. Even if more victims are speaking out now because of higher levels of workforce participation and gender balance, horizontal network structures increase the likelihood of accruing net benefits from doing so.

2. Online Advertising, Algorithmic News Feeds, and Voter Attitudes

In the same fashion, the voter, who maintains no opinion of a candidate in isolation, may change her attitude when confronted by a web of connections, likes, and retweets. Of course the geographic concentration of voting preferences has long existed prior to social media platforms, though this might be characterized as an outcome of shared economic interests vis-à-vis geographic competitors; however, social media platforms, with their sprawling architectures, can amplify existing forms of concentration and create new ones, especially along non-economic dimensions where geographic competition for resources is less important. For instance, various hashtag campaigns related to social policies such as parental rights to raise genderless children, elimination of aggressive policing, and provision of safe spaces for sharing experiences of sexual violence have had broad geographic appeal (despite concentrations in urban centers) and have made at least a marginal impact on state legislatures.[47]

Where geographic concentration of voter preferences discourages support for those policies, social media platforms provide approval benefits across physical distances that can tip support in local referendums and elections. These effects mirror those experienced with the growing syndication of television and radio broadcasting in the past.[48] However, a

---

[47] *See, e.g.*, Tanya Sichynysky, *These 10 Twitter Hashtags Changed the Way We Talk About Social Issues*, THE WASHINGTON POST, (Mar. 21, 2016), https://www.washingtonpost.com/news/the-switch/wp/2016/03/21/these-are-the-10-most-influential-hashtags-in-honor-of-twitters-birthday/?utm_term=.0da730336fa9 (describing the broad appeal and success of various hashtag campaigns).

[48] *See* Adam Candeub, *Media Ownership Regulation, the First Amendment, and Democracy's Future*, 41 UC DAVIS L. REV. 1547, 1587, 1603 (2008) (noting that geographic markets are largely delineated by federal and state law through

key difference is that broadcasting requires significant investment and involves institutional gatekeeping through shareholder pressures and governmental licensing.[49] As a result, the pool of norm entrepreneurs who leverage traditional media to make validity claims over networks is decidedly smaller. While these opinion-makers wield substantial influence, institutional gatekeeping exerts pressure on the contents of their messages.[50] Thus, the political claims made within analog networks are restricted and more tightly controlled. With a narrower set of ideas competing for validity, people place a higher value on the personal qualities of the norm entrepreneur when validating across claims, leading to phenomena such as celebrity newscasters. Social sanctioning, as a result, tends to flow more vertically and less horizontally.

By contrast, norm entrepreneurs who make claims by leveraging social media make use of horizontal and open network structures. While this has created space for a greater number of entrepreneurs and can dilute the strength of messaging, it has simultaneously increased opportunities for aggregating previously untapped approval benefits and has led to greater instances of validation of factual and normative claims (including false ones) which would have been otherwise contained by the geographical dispersion of social approval benefits.[51] But the most

---

spectrum allocation and cable franchise, and discussing the reduction of cross-ownership prohibitions and the explicit rejection of a judicial basis of "viewpoint diversity").

[49] *See* Peter J. Alexander & Keith Brown, *Policy Making and Policy Tradeoffs: Broadcast Media Regulation in the United States*, THE ECONOMIC REGULATION OF BROADCASTING MARKETS 255, 258–59 (noting that media firms face large, up-front fixed costs); *see also* Richard L. Hasen, *Cheap Speech and What It Has Done (To American Democracy)*, 16 FIRST AMEND. L. REV. 200, 205 (2018) (noting that barriers to entry into media have dropped and this has amplified concern for misinformation and propaganda).

[50] Apart from accountability exerted by shareholders and licensors, broadcast communications are highly public and subject to greater levels of scrutiny by the press, fact-checkers, and political opponents. Social media communications, in contrast, are often directly aimed at groups segmented on the basis of private information held by the social media platform. *See* The Honest Ads Act, § 1989 *5 (2017); *see also Disinformation: A Primer in Russian Active Measures and Influence Campaigns, Hearing Before Senate Select Committee on Intelligence*, 115th Cong. 30–41 (2017) (statement of Clint Watts, Robert A. Fox Fellow, Foreign Policy Research Institute) (comparing broadcast news that is difficult to manipulate because it requires actual control of the organs of media with social media news that is easy to manipulate because it requires no control, it is not curated, and is conducive to rapid proliferation).

[51] More precisely, untapped approval benefits were contained by a lower level of

significant difference for existing law is that a larger pool of norm entrepreneurs can now leverage social networks, albeit open and horizontal ones, without being subject to institutional accountability. A reduction in institutional accountability through the avoidance of traditional gatekeepers is additive with a reduction in personal accountability through the ability to remain anonymous.

The Klobuchar, Warner, and McCain proposal essentially mandates disclosure and unmasks the norm entrepreneur. Facebook and Google users are accustomed to seeing their newsfeeds and search results as an organic flow of communications. KWM would alert users to communications that have been disguised as organic but are actually paid political advertisements.[52] It additionally requires the disclosure of who is paying for those advertisements. The underlying premise, which mirrors the existing rationale for regulating offline political advertising, is that people evaluate the credibility of validity claims based upon substance and source. The idea that the source matters is the lynchpin of contemporary ethics in journalism.[53] Sourcing enlarges accountability. By doing so, it guarantees a baseline level of truthfulness in reporting, enhances democratic discourse, and raises the quality of political candidates. In the long run, sourcing nurtures robust political institutions. However, these

---

density of social relationships. To the extent that analog social networks exhibit a high level of density, they too, generate the effects observed over high-density digital social networks. Higher levels of density generate higher levels of social interaction, which lead to marginal increases in social sanctioning and resultant fact and norm proliferation. *See* Yoshinobu Zasu, *Sanctions By Social Norms and the Law: Substitutes or Complements?*, 36 J. L. STUD. 379, 379 (2007) (providing a model where increased density and social interaction leads to higher levels social sanctioning and normative behavior).

[52] In addition to disclaimer requirements, KWM mandates the creation of a public database of online political advertising purchased by advertisers who spend more than $500 per year and sold by social media platforms with more than 50 million unique U.S. visitors per month. The Honest Ads Act, § 1989 *6 (2017). The databased must include a copy of the ad, identification of the target audience, the number of views, the first and last time the ad was displayed, and the name and contact information of the purchaser. *Id.* The purpose is to allow watch-dog organizations to offer near real-time accountability. *See* Yochai Benkler, *Election Advertising Disclosure: Part 1*, HARV. L. REV. BLOG, (Oct. 31, 2017), https://blog.harvardlawreview.org/election-advertising-disclosure-part-1/.

[53] Society of Professional Journalists, Code of Ethics, https://www.spj.org/ethicscode.asp (noting that journalists should "identify sources clearly" and "[c]onsider sources' motives before promising anonymity").

effects assume that people evaluate factual and normative claims on the basis of careful reasoning. Recent studies in social psychology demonstrate that people assess claims, especially political ones, more often as members of a team.[54] Instead of forming political attitudes and beliefs on issues like health care, global warming, and immigration by deliberating facts, people act more like sports fans. When people are rooting for the Cowboys or Patriots, they do not engage in rational deliberation; they are simply expressing loyalty to a team. Paul Bloom explains that "[t]o complain that someone's views on global warming aren't grounded in facts, then, is to miss the point"[55]—political views should be understood "not as articulated conclusions, but rather as 'Yay, team!' and 'Boo, the other guys!'"[56]

Inasmuch as political discourse resembles a team sport more than rational debate, social media platforms cannot manipulate how its users appraise political facts and norms. Increasing institutional and personal accountability can do little to change minds. It seems highly unlikely, for instance, that expectations of a criminal indictment would persuade a Trump supporter to convert to Team Clinton or vice-versa. To be sure, social media platforms can engage in hyper-targeting and enable their advertisers to stoke fears and rouse anger within narrowly segmented groups.[57] But as group membership thins, it becomes increasingly likely that views cannot be changed. This is a direct result of systemic network composition and architecture: within a narrow subnetwork, peers are more likely to confer approval benefits and less likely to confer disapproval costs.[58] Undecided voters, or perhaps independents, are far more likely to

---

[54] *See* Philip M. Fernbach, et al., *Political Extremism Is Supported by an Illusion of Understanding*, 24 PSYCHOL. SCI. 939, 939 (2013).

[55] BLOOM, *supra* note 24 at 236.

[56] *Id*.

[57] *See, e.g.*, Kurt Wagner, *Facebook's Reliance on Software Algorithms Keeps Getting the Company into Trouble*, (RECODE, Sept. 14, 2017, 9:44 PM), https://www.recode.net/2017/9/14/16310512/facebook-mark-zuckerberg-algorithm-ad-targeting-jews (noting that Facebook enabled advertisers to target users based on racist attitudes).

[58] Hyper-targeting of ads and extreme personalization of news generally increases group polarization and reduces meaningful engagement "across the aisle" with ideas. *See* Cass R. Sunstein, *Guest Post: Is Social Media Good or Bad for Democracy?*, FACEBOOK NEWSROOM, (Jan. 22, 2018) https://newsroom.fb.com/news/2018/01/sunstein-democracy/; *see also* Cass R. Sunstein, *The Law of Group Polarization*, 10 J. POL. PHIL. 175, 175 (2002). However, it is unclear whether observations of greater polarization are the result of disinhibition toward revelation of true preferences in the context of a safe and approving atmosphere.

be persuaded by political messaging, but this group is simultaneously far less likely to be algorithmically identified as a worthy recipient of a dubious ad. Moreover, undecided voters are more likely to engage in deliberate reasoning as opposed to political team sports; they consequently view dubious ads with higher levels of suspicion. Note that ongoing maintenance of an advertisements database, such as the one envisaged by the KWM proposal, may reduce instances of hit-and-run political advertising, but do little towards increasing the quality of political discourse. Polarized groups will simply debunk or vindicate the archived advertisements according to their preexisting tastes. On the other hand, if debunking will reduce a tendency to view untruthful advertisements, independent of preexisting tastes, then the database may be helpful toward reducing the impact of inflammatory ads ex post and their creation in the first place.[59] Either way, effective regulation should target the composition of teams, especially if the social psychology research is correct.

One approach could be to limit the types of groups that may be algorithmically identified and available for impression. As the target group widens, the likelihood increases that patterns of approval will be checked with patterns of disapproval, and that false claims will fizzle out. By backward induction then, the number of weakened claims will decrease. This approach has several design advantages. First, it side-steps the argument sometimes raised by social media platforms that online advertising is too short to include a full disclaimer.[60] Second, it is impervious to botnets, sockpuppets, and other types of synthetic social

---

[59] This is because content producers would have no incentive to produce ads that generate no income. *See* Samanth Subramanian, *Inside the Macedonian Fake News Complex*, VOX, (Feb. 15, 2017), https://www.wired.com/2017/02/veles-macedonia-fake-news/ (noting that the income is derived from impressions). While the reduction of inflammatory ads may have no impact on persuading core supporters, it can, over time, increase the overall quality of political discourse if reduction leads to greater levels of rational deliberation and truth-seeking. *See* Benkler, *supra* note 52 (noting that an effective database could keep campaigns more honest and constrained, by allowing internet users "from professional journalists and nonprofits to concerned citizens with a knack for data, to see what the campaigns and others are doing, and to be able to report on these practices in near-real time to offer us, as a society … the ability to understand who, more generally, is trying to manipulate public opinion and how").

[60] *See* Mark Zuckerberg, FACEBOOK, (Sept. 21, 2017), https://www.facebook.com/zuck/posts/10104052907253171 (promising that going forward, Facebook will disclose the identity of the advertiser and that it will provide a link to a page displaying all of the advertisements a particular advertiser is running to any audience on Facebook).

behavioral marketing that are more or less resistant to the effects of disclaimer requirements.

Finally, it avoids chilling speech. Critics of KWM and similar proposals argue that disclosure regulation flouts First Amendment principles and will stifle speech.[61] By shifting regulatory focus away from the speaker and toward questionable forms of user segmentation, regulation of online advertising would track existing rules against racial and other forms of group profiling. Admittedly, in commercial contexts, these rules have been primarily applied to economic actors who, among other things, deny services to members of a protected class.[62] A narrowly tailored systemic prohibition against segmenting users into hate groups and offering advertisers access to them would likely survive rational scrutiny. Enhancing the quality of political discourse or minimizing outside interference in electoral processes are legitimate government interests; and it is especially difficult to classify the groups, or their advertisers, as protected classes. In any case, Facebook has already implemented internal rules against targeting racist groups identified by its algorithm, though it remains unclear how the rules will be applied over time. It is important to highlight what, exactly, the rules do: they essentially subject advertisements that would experience net approval within a targeted subnetwork to a broader group of people who disapprove.[63] Instead of suppressing platform speakers, they configure

---

[61] *See*, *e.g.*, Eric Wang, *Analysis of Klobuchar-Warner-McCain Internet Ads Legislation (S. 1989, 115th Cong.)*, 2017 INST. FREE SPEECH 1 (asserting that legislation which attempts to limit foreign influence by broadly regulating free speech will burden online political speech).

[62] *See* United States v. Carolene Products Co., 304 U.S. 144, 153 n.4 (1938) (noting that laws which impact ordinary commerce, including freedom of contract, are unconstitutional unless they rest upon a rational basis; while economic regulations that "prejudice . . . discrete and insular minorities" require greater scrutiny when the prejudice "tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities").

[63] Note that this approach targets the demand-side, that is, the consumers of political advertisements, and can be supported by additional efforts to combat computational propaganda directed at the supply-side such as: requiring users to identify themselves to the platform and authenticate their accounts before being permitted to post publicly anonymous content; limiting the number of posts that can be made by a single account within a specified time frame; and using human verification systems like (CAPTCHA) to combat automated messaging. *See Terrorism and Social Media: #Is Big Tech Doing Enough?: Hearing Before U.S. Sen. Comm. on Commerce, Science, and Transportation*, 115 Cong. 3 (2018) (statement of Clint Watts, Robert A. Fox Fellow, Foreign Policy Research Institute).

platform architecture more widely to generate a mixed network composition, which reduces the incentive to produce inflammatory speech in the first place.

1. Social Media and Campus Speech

Once a norm entrepreneur makes a validity claim that has persuaded at least one person, that person acts as an enforcer of the entrepreneur's opinion or norm to the extent that she publicizes her approval to others. She can also enforce contrary opinions and norms by publicizing her disapproval. Oftentimes, enforcement is costly. The persuaded individual, who is unsure of the reaction that her social approval or disapproval may generate, will fear a negative response from members of her social network. She may fear backlash and disapproval for being preachy or for taking a position that she erroneously believes to be consistent with her network. Or enforcement may be costly simply for the time it takes to write an opinionated message to a friend, police a lengthy Twitter account or Facebook News Feed with likes, or bother with an apathetic acquaintance or unknown stranger. To be sure, technology has reduced the economic costs of social enforcement dramatically. No longer does one need to draft a letter to the editor, appear at the town square for an evening *passeggiata*, or travel door-to-door with a petition to publicize one's views and change social attitudes. At the same time, a tweet, a like, or the casual retweet can carry less opprobrium or validation than the lengthier face-to-face exchanges of yesteryear, though not always. Today's social networks are large, provide immediate feedback, and can appear downright sincere when its participants broadcast intimate personal details.[64] These features, among others, generate meaningful second-party enforcement that shape social attitudes and social norm compliance. As the sanctioning strength of a social network increases, taking a position on behalf of a norm entrepreneur becomes costlier, not just in terms of time

[64] *See* Anna M. Lomanowska & Matthieu J. Guitton, *Online Intimacy and Well-Being in the Digital Age*, 4 INTERNET INTERVENTIONS 138, 139 (2017) (noting that online interactions can mirror levels of intimacy of offline interactions, and that in some contexts, "can actually accelerate intimacy formation in comparison"). Consider that pop-stars Erykah Badu and Jay Electronica tweeted 4595 "protected" followers, that is, only those Twitter users who have chosen to follow them (as opposed to any Twitter user), live descriptions of the delivery of their baby. *See* Jayson Rodriguez, *Erykah Badu, Jay Electronica Blog Child's Birth in Real Time on Twitter*, MTV NEWS, (Feb. 2, 2009), http://www.mtv.com/news/1604057/erykah-badu-jay-electronica-blog-childs-birth-in-real-time-on-twitter/. As Lomanowska and Guitton emphasize, sharing intimate personal experiences such as childbirth increases the sincerity of online interactions. *See* Lomanowska & Guitton, *supra* note 64 at 139.

and effort, but also for the negative response it may produce. Policing a validity claim may elicit a net positive response as well. Networked acquaintances who approve of supporting or opposing a claim provide increased utility and satisfaction to the enforcer.[65] Inasmuch as the benefit from receiving aligned social responses exceeds the costs of eliciting them, people will continue to engage in second-party enforcement.

Seeking positive responses and avoiding negative ones leads a person to choose like-minded acquaintances and generally select a network composition that reflects her personal beliefs. Note that this network-selection behavior is entirely rational and helps explain why social networks systemically exhibit high levels of herding and polarization, and why they tend to aggressively reinforce existing patterns of belief. In addition to network selection, contemporary platform architecture dramatically limits the disapproval costs that users face. Social media platforms steer users toward upvotes and likes (and away from downvotes and dislikes), and obscure information about being unfriended or blocked by another.[66]

Historically, universities have fostered political speech partly for the values and ideas that flourish there, but also because there exists a density of social interactions amongst students who share those values and ideas. Students who protest for or against providing a platform to a visiting speaker incur few costs and receive nearly certain benefits from their respective social networks. While university speech regulations concerning time, manner, and place limit speech activity so as to minimize conflict among the university body and interference with university responsibilities,[67] this policy implies that threat of disruptive protest is

---

[65] *See* SMITH, *supra* note 29 at II.iii.2.1 (noting that the love of praise and the dread of blame motivates action).

[66] For instance, Facebook users can only unlike comments, photos, or posts that they have previously liked. *See* How Do I Unlike Something?, FACEBOOK, https://www.facebook.com/help/226926007324633?helpref=uf_permalink (last visited Mar. 30, 2018). Because of the time and complication associated with discovering who has unfriended or blocked them on Instagram, Facebook, and Twitter, interested users typically resort to third-party apps. *See* Joe McGauley, *How to See All the Jerks Who Unfriended You on Facebook*, THRILLLIST, (Dec. 23, 2016, 2:46 PM), https://www.thrillist.com/tech/nation/how-to-see-who-unfollowed-you-tracking-friends-on-instagram-facebook-and-twitter.

[67] *See*, *e.g.*, Berkeley Campus Regulations Implementing University Policies, Section 300, Regulations Concerning Time, Place, and Manner of Public Expression, BERKELEY – UNIV. OF CA, http://sa.berkeley.edu/uga/regs., (last updated Aug. 23, 2011) (stating that its regulations concerning time, place, and manner of public expression are designed to prevent interference with the

increasingly successful the more likely it disrupts campus life. Moreover, disruption provides increasingly large social approval benefits to its members as a social network increases in size and narrows in viewpoint.[68] Simultaneously, counter-protestors who organize for continuance of the speech experience inverse benefits and costs. Given the zero-sum nature of campus conflict, net social benefits are a function of the size of the two opposed groups, and efficiency—at least in the short run—generally favors the preference of the larger group.[69]

To the extent that protest generates conflict, the students externalize a number of costs that are incurred by the university. The most prominent are monitoring and security costs. Ideally, these would tend toward zero, which explains why norm entrepreneurs and lawmakers invest in norms that establish procedural excellence and promote rational deliberation within fora.[70] On the other hand, there has been a turn in social attitudes amongst student bodies, which is reinforced by heightened social network density through the use of social media, that sound university procedure must prevent the circulation of hateful substance.[71] Insofar as this approach generates lasting social benefits, it may be efficient over time. Even so, if taken to its logical limits, universities can no longer exclusively rely upon time, manner, and place restrictions for making determinations of permissible speech. Given a university policy to prohibit disruptive speech, such restrictions will only continue to have bite on campuses inasmuch as the students themselves embrace the speech as substantively valid and refrain from disruption.[72] Again, it bears emphasis

---

university's conduct and affairs).

[68] Disapproval costs from taking a contrary view increase in addition.

[69] Group size may matter little for long-run efficiency if the views of the majority lead to a residual social benefit or cost

[70] *Cf.* RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 599–600 (5th ed. 1998) (noting that the economic goals of civil and criminal procedure are to minimize errors of judgment and the costs of administering law). For an early modern historical account of the limitations of discursive excellence toward solving social problems, see TERESA M. BEJAN, MERE CIVILITY: DISAGREEMENT AND THE LIMITS OF TOLERATION *passim* (2017).

[71] *See* ERWIN CHERMERINSKY & HOWARD GILLMAN, FREE SPEECH ON CAMPUS 13 (2017) (noting that in contemporary campus speech issues, "it is the students who demand that the campus take action against speech they find offensive").

[72] To remain consistent with the First Amendment, schools must be careful to base their decisions to prohibit speech on disruption of student learning and school environments and not on the offensive character of speech contents. *See* Kowalski v. Berkeley Cty. Sch., 652 F.3d 565, 574 (4th Cir. 2011), (ruling a MySpace chat group which encouraged vulgar and offensive comments about another student not protected because the distress it inflicted caused school

that granting de facto adjudicatory power to the students may be socially efficient in the short run if a majority disfavors the speech. In the long run, efficiency requires that disruption lead to a lasting social benefit.[73]

One clear area for university action is to curb norm entrepreneur activity that provides few social benefits to students and accomplishes few long-run social goals, and is instead aimed at increasing monitoring and security costs incurred by the university and degrading procedural norms. When unfriendly outsiders organize a protest, and perhaps a counter-protest at the same time, their agitation efforts are unambiguously meant to drive up institutional costs and reduce the quality of discourse. As recently seen in Texas, foreign agents were able to organize two-sided protests through Facebook advertising for $200.[74] If we loosely delineate

---

disruption), *cert denied*, 565 U.S. 1173 (2012); J.S. ex rel. Snyder v. Blue Mountain Sch. Dist., 650 F.3d 915, 928 (3d Cir. 2011), (finding a phony MySpace profile created by student to ridicule school principal protected under the First Amendment because the spoof profile did not substantially disrupt student learning and school environment), *cert denied*, 565 U.S. 1156 (2012); Padgett v. Auburn Univ., Case No. 3:17-CV-231-WKW (M.D. Ala. 2017) (protecting speech because university prohibition was unlawfully based upon its offensive content); *see also* Erwin Chermerinsky, *Hate Speech is Protected Free Speech, Even on College Campuses*, VOX (Dec. 26, 2017), https://www.vox.com/the-big-idea/2017/10/25/16524832/campus-free-speech-first-amendment-protest (referencing the Padgett case and noting that the Supreme Court has consistently held that public institutions, including universities, cannot prohibit speech on the basis that it is deeply offensive); *see also* Robert C. Post, *There Is No 1st Amendment Right to Speak On a College Campus*, VOX (Dec. 31, 2017), https://www.vox.com/the-big-idea/2017/10/25/16526442/first-amendment-college-campuses-milo-spencer-protests (noting that "[t]he limits on the university's ability to regulate the speech of its students are . . . demarcated by the limits of its educational reach over students").

[73] An analysis would track basic First Amendment policy where the chilling of speech is weighed against the probabilistic outcome that the speech creates an even greater social loss. *See* United States v. Dennis, 183 F.2d 201, 212 (2d Cir. 1950) (J. Learned Hand) (explaining that courts must in each case "ask whether the gravity of the 'evil', discounted by its improbability, justifies such invasion of free speech as is necessary to avoid the danger"), *aff'd*, 341 U.S. 494 (1951). *See also infra* § IV.A.

[74] Though this example did not occur on a campus, it is instructive. *See* Natasha Bertrand, *Russia Organized Two Sides of a Texas Protest and Encouraged 'Both Sides to Battle in the Streets'*, BUSINESS INSIDER, (Nov. 1, 2017, 1:25 PM), http://www.businessinsider.com/russia-trolls-senate-intelligence-committee-hearing-2017-11 (reporting Senate Intelligence Committee Chairman Burr estimating costs of agitation at $200).

two social networks, one for the protestors, and one for the counter-protestors, each member's welfare is obviously enhanced by increased approval benefits and avoidance of disapproval costs by participating. Protest participants receive nearly certain social approval benefits within their respective social networks, but these tend to be outweighed by externalized security costs and manufactured institutional decay. Only if the two networks were systemically unified would there be opportunities for simultaneous approval and disapproval, which would, in turn, reduce the likelihood of the occurrence of the costly protest and counter-protest. This means that university policies which foster the creation of plural network membership reduce susceptibility to institutional attack. For instance, political advertising could be confined to a campus-wide Facebook or Kialo page.[75] By restricting political advertising to a broadly viewed page, the university can drive convergence of approval and disapproval within a singular platform location to the extent that its student body is relatively balanced with multiple views.[76] Guaranteeing that the advertisements are targeted broadly to the entire university body may be costly but perhaps not prohibitive. Of course to the extent that Facebook and other platforms prohibit advertisers from targeting hate groups as suggested in Section II.B.2, the probability of agitation within isolated social media platform locations, and the consequent increases in monitoring and security costs incurred by campuses, decrease.

III. SOCIAL NORMS AND THE SELF

A. The Emergence of Guilt and Pride

Because social networks provide approval and disapproval benefits and costs, they can change attitudes, calibrate beliefs, and make compliance with newly created norms (or old ones) go viral. In addition, they can generate compliance with legal rules. Rules that are rarely enforced by law, because of problems with detection or because the costs of bringing a claim are greater than the relief sought, can be enforced with

---

[75] For an overview of Kialo, a social media platform that splits arguments into binary "for" and "against" trees to enable apprehension of opposing views and encourages users to rank arguments on the basis of their reasoned qualities, see Jonathan Margolis, *Meet the Startup that Wants to Sell You Civilized Debate*, FIN. TIMES, Jan. 24, 2018, https://www.ft.com/content/4c19005c-ff5f-11e7-9e12-af73e8db3c71.

[76] Efforts toward curbing external costs might be additionally supported through the creation of content aimed at reducing hate speech through programs like "Creators of Change". *See* Creators of Change, YOUTUBE, https://www.youtube.com/yt/creators-for-change/ (last visited Mar. 30, 2018).

social sanctions.[77] The non-recycler who routinely ignores local regulations may begin to recycle when confronted by neighbors; or the parent may attempt to control a troublesome child more in public than at home. And to the extent that a would-be serial harasser is a member of a robust network that disapproves, he too, is more likely refrain from asocial misconduct. None of this is surprising or novel, but it lays the foundation for what happens next, when sanctions of approval and disapproval enforced by others transform into sanctions of personal guilt and pride enforced deliberately by one's self.

Carrying out social sanctioning can be costly for the enforcer. While social media platforms have dramatically reduced the necessary time and effort, enforcers can, nonetheless, experience meaningful costs when their sanctioning backfires and generates negative reactions throughout their network even though these are limited by platform architecture.[78] Enforcers sanction then, when the expected value of sanctioning is positive. They weigh the benefits, which are composed of some measure of how deeply they care about the propagation of a social attitude or norm, and the expected net approval benefits that accrue to them when they disseminate it. Because net approval benefits are reduced by any backlash disapproval that the enforcer thinks could be forthcoming from her network, second-party social sanctioning tends to peter out.

When people come into contact with a validity claim that they accept as true, they can develop sources of internal sanctioning that are strictly independent of other people's beliefs.[79] For instance, a person who accepts the claim that the Earth is warming due to human activity can feel internal pride from accurately recycling. She may carefully sort the plastic from the paper not because she is under the watchful gaze of a stranger or friends, but because she internally reflects upon her beliefs and experiences emotional benefits from taking a consistent action. Once she posts a selfie, seated beside two piles of plastic and paper, she is positioning herself to receive probabilistic approval benefits from her network. If she is running late, and simply disposes of the recycling unsorted, she experiences internal guilt. A friend who catches her in the act, and posts a photo on Instagram, may generate disapproval costs inasmuch as her network disapproves. Two points are of interest. First, her

---

[77] *See* Cooter, *supra* note 22 at 1597 (noting that intimate relationships are a primary influence on a person's character); Zasu, *supra* note 51 at 379 (noting that informal sanctions can substitute for law).

[78] *See supra* note 66.

[79] *See* SMITH, *supra* note 29 at Part I.iii.1 (noting that systems of internal approbation and self-love are based on reason and sentiment).

internal guilt or pride for accurately recycling is the result of earlier work completed by a norm entrepreneur. Second, once a claim is internalized, internal sanctioning is not probabilistic because it does not depend upon the actions of others. While approval and disapproval may or may not be forthcoming given prevailing attitudes and the composition of her network, her internal guilt and pride is certain.[80] This point matters. It means that enforcers of validity claims can reliably economize on costs.

Even when facts and norms are widely accepted and sanctioning is unlikely to trigger backlash, enforcement still requires time and effort. More importantly, enforcement of broadly accepted norms within a network can generate disapproval because it can signal that the norm has not been fully internalized by the speaker. For instance, explicitly disapproving of a norm against holding out religious beliefs in a commercial setting on a particular message board might actually signal that the enforcer considers doing so an option.[81] Other complexities related to the sincerity of the sanction, such as the enforcer's timing and context, may generate disapproval as well. Sounding insincere, preachy, or out of touch, carries a cost. Thus, even though expected backlash may approach zero as the fact or norm becomes more widely held, there is always a non-zero probability of disapproval. For this reason, levels of second-party social sanctioning fade, not so much for changing tastes or the reduced salience of a once important issue, but rather for the emergence of lower-cost enforcement via guilt and pride. Recycling still matters, but enforcers can rely on sources of internal guilt and pride and simultaneously avoid appearances of insincerity, arrogance, or ignorance.

It should be clear that the deterioration of approval and disapproval may not occur for every social attitude and norm. When internalization of a claim is slow or non-existent over a network, then social enforcement will generate net benefits for enforcers who care deeply about a fact or norm (so long as expected costs are sufficiently low). In those cases, patterns of approval and disapproval will subsist, but it bears emphasis, that in other cases, it makes economic sense for patterns

---

[80] This is true so long as the norm prevails internally within the individual. If a norm entrepreneur makes a new validity claim, which unseats the underlying fact or norm that is driving an individual's pride and guilt, then the previous norm no longer prevails and guilt and pride sanctioning will lapse. However, it is incorrect to say that a fixed and prevailing norm that has been internalized generates expected values of guilt and pride. So long as the norm is internalized, guilt and pride are certain by definition.

[81] *Cf.* SETH STEPHENS-DAVIDOWITZ, EVERYBODY LIES 11–12 (2018) (highlighting the complications of uncovering true preferences on the basis of search results).

to dissolve. At dissolution, the underlying norm remains supported by internal sanctioning and continues to exert compliance and attitudinal effects.

B. The Emergence of Unconscious Compliance

As facts and norms become more deeply internalized, rational individuals can maximize their return from holding beliefs and complying with norms by forgetting that they once required first-party enforcement to conform. A person who gives up smoking may feel a sense of pride from declining a cigarette within the first few years of quitting, when the chemical and psychological urges are still strong. But a former smoker who declines a cigarette a decade later experiences less pride if any at all. Economists would say that the marginal benefits of pride decline over time.[82] Moreover, engaging in pride (or guilt) to generate compliance or conformity with a social attitude may itself be costly inasmuch as it recalls an earlier calculus that tolerated non-conformity as an option. A former smoker who declines a cigarette while pregnant for instance, may experience guilt for recalling that she once considered smoking while pregnant an option, and may avoid rewarding herself with pride. This is not to say that all norms and attitudes eventually become unhinged from first-party enforcement. The point is simply that sometimes, forgetting that one was guilty or prideful can be a rational decision aimed at maximizing the benefits of internalization.[83] Over time, this form of psychological evolution can, in some instances, generate involuntary and unconscious compliance with normative behaviors (and apparent acts of self-sacrifice), which generate no meaningful benefits for the actor.

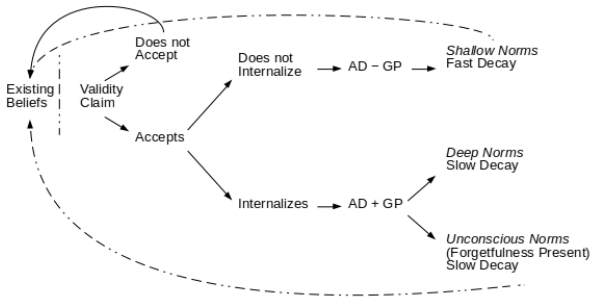

Figure 2: How Social Norms Go Viral[84]

---

[82] *See* BECKER, *supra* note 21 at 50.

[83] *Id.*; *see also* Gary S. Becker, *Habits, Addictions, and Traditions*, 45 KLYKLOS 327, 327 (1992).

[84] Approval, disapproval, guilt, and pride are denoted A, D, G, and P.

It is important to keep in mind, however, that unconscious behavior and attitudes can be set in motion by rational maximizing norm entrepreneurs. Consistent with their financial and social goals, norm entrepreneurs invest in validity claims to the extent that they expect a positive return on their claim-making investments.[85] This instrumental account of norm creation and evolution can be used to help explain a number of phenomena, including differences in attitudes toward immigration, climate change, and church-state separation. Before turning to these examples, it is useful to sketch a basic three-part taxonomy. It may be said that shallow norms operate on the surface of social networks. They subsist primarily on the basis of approval and disapproval and tend to decay at a relatively high rate. As a result, they generate lower returns. Deep norms rely jointly on external approval and disapproval as well as internal pride and guilt. Because these norms have socially matured, they often decay at lower rates and generate higher returns to the norm entrepreneur. Unconscious norms operate on the basis of rational forgetfulness and are rarely conscientiously challenged. They decay at the highest rate, and require the highest level of investment to dislodge.

1. Immigration

Social attitudes toward immigration are shaped by norm entrepreneurs and second-party enforcers. There are a number of instrumental reasons why entrepreneurs and enforcers may want to shape societal views toward immigration. They may believe that diverse societies hold the greatest prospect for lasting peace and human progress; they could be concerned with international competitiveness in the face of declining population growth within their nation; perhaps they may represent a narrow set of interests that prefer higher or lower levels of competition for jobs amongst workers—both skilled and unskilled; or they may have internalized a generosity norm in the past and, as a result, benefit from professing compassion to strangers. In all likelihood, a desire to express an attitude toward immigration is based upon some combination of reasons, these and others, and has been developed with great nuance.

Given the salience of validity claims made in relation to immigration in contemporary political discourse and communication, norms and attitudes are clearly operating at the shallow or deep level. To the extent that each person in a society feels guilt or pride for holding pro-

[85] For social media platforms, returns include profit, corporate image, long-term viability, good citizenship, and a friendly regulatory environment. *See infra* § IV.C.

or anti-immigration beliefs, the attitude can be said to be operating at a deep level. Norm entrepreneurs who seek to dislodge and replace that attitude must invest in claims tailored toward generating guilt and pride throughout a social system. For instance, pro-immigration messaging can be anchored in common-cause: The United States is an immigrant nation, its people are immigrants, and individuals who disapprove of immigration today have reason to feel guilty since their forebears relied, in part, on pro-immigration attitudes to drive policy. These types of claims are more difficult to make in non-immigrant nations. For example, in Japan, anti-immigration attitudes can be traced to one of its earliest recorded poems, which celebrated an "eightfold fence" that separated it from other lands and peoples.[86] To the extent that norm entrepreneurs seek to shape prevailing attitudes toward immigration within Japanese society, they have to appeal to a broader cosmopolitan history that exists outside of Japan. Claims may go viral on the surface of a Japanese social network, on the basis of peer approval and disapproval, but do little toward changing deeply held attitudes and beliefs. In this case, greater levels of investment in claims are required for generating change.

When the President uses profane language to describe a particular set of countries and expresses disdain for their immigrants,[87] it does little to change existing attitudes toward immigration policy to the extent that American attitudes and norms operate on the surface, and are primarily shaped by second-party social sanctioning anchored in peer-approval and disapproval. In this setting, reactions to inflammatory statements that target the speaker and generate patterns of social sanctioning around immigration are based more upon a person's alignment with the speaker and less upon the policy contents of the message. Contemporary rational deliberation, with its characteristic dryness and dearth of provocative rhetoric, is a relatively weak player in the world of political team sports. On the other hand, internalization, at least of political attitudes and beliefs, requires greater levels of personal reflection for which basic social media platforms seem poorly equipped to systemically generate.[88]

---

[86] The poem is found in the imperial anthology, *Kojiki*, dated to the early eighth century. 古事記 (KOJIKI) [RECORDS OF ANCIENT MATTERS], 396–402 (trans. Basil Hall Chamberlain, 2d ed., 1932) (c. 712).

[87] *See* Julie H. Davis, et al., *Trump Alarms Lawmakers With Disparaging Words for Haiti and Africa*, N.Y. TIMES, (Jan. 11, 2018), https://www.nytimes.com/2018/01/11/us/politics/trump-shithole-countries.html.

[88] This may be because they engender terse, normative messages that contain less content, analysis, and factual support. On the other hand, shorter messages may be easier to process. *See* LARRY SAMUELSON, EVOLUTIONARY GAMES AND EQUILIBRIUM SELECTION 24 (1998) (noting that apprehension increases with

It may be that prevailing American social attitudes toward immigration are primarily based upon social approval and disapproval amongst acquaintances and have little to do with internal values. For instance, a voter may feel little pride (or guilt) for supporting (or opposing) an altruistic policy that increases immigration levels from conflict- and disaster-stricken countries or countries with low average incomes. Any guilt, if present, may be outweighed by one's identity with a chosen political team. In this case, a viral norm will quickly decay throughout a social network.

Recall that social media platforms generally increase the density of social interactions.[89] What does this mean for immigration? If immigration attitudes are primarily based upon internal values, and social media messaging tends to target external patterns of group approval and disapproval, then immigration discourse will remain relatively unperturbed. Norm entrepreneurs must make claims that target internal guilt and pride, and these types of claims are less likely to travel through media such as casual tweets, comments, and upvotes. Only if deep claims aimed at internalization take root, will guilt and pride increase, and will social norms and attitudes become more consistent over time.

1. Climate Change

This point can be clearly seen in current disagreements over why Americans view climate change with more skepticism than Europeans. Cass Sunstein has suggested that political support for combating terrorism is greater than support for climate change because the former is more salient.[90] People can see the effects of terrorism and imagine themselves harmed. They cannot do the same for climate change. Public attitude surveys seem to support this claim. Latin America and Africa, two regions that have experienced relatively high levels of drought, are more concerned with climate change than other regions.[91] Likewise, Australia, which experiences regular forest fires, maintains strong support for public

---

"sufficiently simple" messaging that commands attention).

[89] *See supra* note 51.

[90] Cass R. Sunstein, *The Availability Heuristic, Intuitive Cost-Benefit Analysis, and Climate Change*, 77 CLIMATE CHANGE 195, 195 (2006).

[91] Bruce Stokes, Richard Wike, & Jill Carle, *Global Concern About Climate Change, Broad Support for Limiting Emissions*, PEW RESEARCH CTR: GLOBAL ATTITUDES & TRENDS, (Nov. 5, 2015), http://www.pewglobal.org/2015/11/05/global-concern-about-climate-change-broad-support-for-limiting-emissions/ (finding 74% and 61% of survey respondents in Latin America and Africa very concerned, and only 54% and 45% of European and American respondents very concerned).

action.[92] Salience cannot fully explain the difference in attitudes between the United States and Europe, however. Both regions are situated in the Northern Hemisphere and have been insulated from dramatic and visible climatic changes, but differences in concern between its citizens vary substantially.[93]

Studies often explain these differences in terms of scientific literacy.[94] Europeans are more knowledgeable in science, the argument goes, and are consequently more concerned with global warming.[95] Some believe that more concern translates into greater levels of second-party enforcement. For instance, a common view is that "there is a fair bit of social pressure to behave in an environmentally responsible manner in places like Sweden."[96] In order to change American attitudes toward climate change then, norm entrepreneurs must invest in claims that increase scientific literacy. On the other hand, more recent studies emphasize the role of political affiliation and ideology, and build on

---

[92] The Climate Institute, *Climate of the Nation 2016: Australian Attitudes on Climate Change*, http://www.climateinstitute.org.au/verve/_resources/COTN_2016_Executive_Summary.pdf (finding that 77% of Australians believe climate change is occurring and continues to grow and that a majority believes the government should do something about it).

[93] *See* Stokes et al., *supra* note 91.

[94] Anthony Leiserowitz, et al., *Climate Change in the American Mind*, Yale Program on Climate Change Communication and George Mason University Center for Climate Change Communication 9–10 (2017) (finding that only 13% of Americans understand that there is a scientific consensus that humans cause global warming); van der Linden et al., *The Scientific Consensus on Climate Change as a Gateway Belief: Experimental Evidence*, 10 PLOS ONE 1, 7 (2015) (finding that beliefs about scientific consensus shape attitudes toward combatting climate change); *cf.* Elke C. Weber & Paul C. Stern, *Public Understanding of Climate Change in the United States*, 66 AM. PSYCHOLOGIST 315, 315 (2011) (noting the difficulty in understanding climate change results in polarization of beliefs).

[95] Jing Shi, *Knowledge as a Driver of Public Perceptions About Climate Change Reassessed*, 6 NATURE CLIMATE CHANGE 759, 759 (2016) (noting survey evidence of greater scientific literacy, and greater concern for climate change, in Europe versus the United States). *But see*, Dan Kahan et al., *The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks*, 2 NATURE CLIMATE CHANGE 732, 723 (2015) (noting that scientific literacy tends to polarize views because people with high literacy use it to retain and justify preexisting beliefs).

[96] Elisabeth Rosenthal, *What Makes Europe Greener than the U.S.?*, YALE ENV. 360 (Sept. 28, 2009), http://e360.yale.edu/features/what_makes_europe_greener_than_the_us.

psychological research that views politics as a team sport.[97] The tension within this research should be clear insofar as increasing scientific literacy generally reduces politicization. Literacy engenders rational deliberation, which decreases the tendency for norms to operate exclusively on the surface of social networks via approval and disapproval. If social psychologists are correct, policy should be directed toward depoliticizing climate change issues, with care being paid to not inflame beliefs grounded in ideology.[98]

It would seem that this can be accomplished at least two ways: increasing scientific literacy and decreasing politicized validity claims made over social networks. What remains unclear, however, is whether first- and second-party enforcement crowd each other out. In other words, will limiting instances of "Yay skeptics, and boo scientists" (and vice-versa) free up space for rational thought and the emergence of guilt and pride? If so, then policy directed toward *reducing* casual social media communications related to climate change may, counterintuitively, encourage the development of deep social norms inasmuch as platform messaging fails to generate internal reflection and simply reinforces existing polarized ideologies with patterns of peer approval.

One systemic approach is to encourage norm entrepreneurs and social network members to make and enforce validity claims related to climate change on platforms or sub-platforms that are dedicated to town-hall-style discourse or scientific debate. In other words, modify the platform architecture for claims that are meant to target internalization. This approach, taken by Reddit, StackExchange and others, has led to a relatively higher quality of idea exchange and has decreased instances of ideological herding and trolling.[99] The platform permits others to reward

---

[97] *See, e.g.*, Kelly S. Fielding & Matthew J. Hornsey, *A Social Identity Analysis of Climate Change and Environmental Attitudes and Behaviors: Insights and Opportunities*, 7 FRONT. PSCYH. 121, 121 (2016) (noting that tensions between political conservatives and progressives drive differences in attitudes toward climate change).

[98] In other work, I have noted that lawmakers should work toward depoliticizing climate science and other gridlocked policies by enacting contingent rules that become effective only if certain scientific facts obtain. *See* Frank Fagan, *Legal Cycles and Stabilization Rules*, THE TIMING OF LAWMAKING 11, 16–18 (Frank Fagan & Saul Levmore eds. 2017); Frank Fagan, *Political Paralysis and Timing Rules*, 91 N.Y.U. L. REV. ONLINE 43, 48–50 (2016).

[99] *See, e.g., How Does StackExchange Stimulate Honest, Open Discourse While Limiting the Effects of Trolling?*, META STACKEXCHANGE, https://meta.stackexchange.com/questions/289629/how-does-stack-exchange-stimulate-honest-open-discourse-while-limiting-the-effe (last visited Apr. 6,

the poster for exceptional participation—not so much for the content that they provide, but for the manner in which they provide it. Yet another systemic approach is to provide links to related content which articulates counter-arguments or places similar arguments within different contexts.[100] By creating an atmosphere that encourages rational deliberation, climate change discourse can be funneled toward social media platform locations that foster the development of deep norms.

2. Religious Expression (and Wedding Cake)

As a norm moves deeper toward unconscious cognitive process, either because its roots are rationally forgotten or simply lost over time, the rationale for following its mandate becomes increasingly difficult to ascertain. Older rationales may evolve and merge with newer ones, be entirely replaced, or become confused, incoherent, or unspoken. Chaotic evolutions are often on display when two opposing norms conflict. Consider that in most Western societies today, there exists a deep norm of free expression of private religious beliefs. The norm is at least as old as the Reformation, if not older, and has been codified in many statutes and constitutions across Europe and the United States.[101] Within many countries, especially Protestant ones, free expression was eventually interpreted as the freedom to choose one's own religious practices. In order to preserve free choice for everyone, it was necessary that one's personal practices did not apply to others. This led free religious expression to engender a norm of separation between church and state.[102] Religious freedom is rarely understood as a normative precursor for

---

2018).

[100] For example, Facebook is testing the integration of its "Related Articles" program into Facebook News Feed. *See* Sara Su, *New Test With Related Article*, FACEBOOK (Apr. 25, 2017), https://newsroom.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles/. Similarly, Apps like Read Across the Aisle nudge users toward articles that are less consistent with their political beliefs. *See* Richard Bilton, *A News App Aims to Burst Filter Bubbles by Nudging Readers Toward a More "Balanced" Media Diet*, NEIMANLAB (Mar. 9, 2017), http://www.niemanlab.org/2017/03/a-news-app-aims-to-burst-filter-bubbles-by-nudging-readers-toward-a-more-balanced-media-diet/. See also note 75 for a description of Kialo, another platform that encourages evaluation of argument and counter-argument.

[101] *See, e.g.*, U.S. CONST. amend. 1; FRENCH DECLARATION OF HUMAN AND CIVIC RIGHTS, Art. 10–11 (1789).

[102] *See* BRAD S. GREGORY, REBEL IN THE RANKS 218, 255–56 (2017) (noting that separation of church and state "continues to be paradoxically enabled by the freedom of religion, which itself was conceived as a solution to problems inherited from the Reformation era").

church-state separation because it presents a partial paradox: if separation leads to secularization of public life, then freedom of expression is truncated to the extent that expression takes place in public. The incoherence will eventually subside when the older religious expression norm is forgotten, discarded, or reconfigured.

The *Masterpiece Cakeshop* case can be understood as a step toward reconfiguration, or at least an emphasis and reaffirmation of the importance that religious expression is free insomuch as it does not impact others in public life.[103] Assuming that the creation of the cake is an expression of the baker's religious beliefs, the benefit that he receives (his dignity interest) is dependent upon denying the couple service.[104] When denied service, the couple incurs a cost (equivalent to their dignity interest). On the other hand, when compelled to create the cake, the couple receives their dignity interest, but the baker is denied his. The case might be easily decided if the magnitude of the parties' dignity were observable and law were content to ignore the interests of society at large.[105] Without an ability to measure the parties' subjective valuations of dignity, law might consider a proxy in the form of the income the baker gives up by denying the couple a cake, and perhaps the economic value of a free cake offered to the couple by another supportive baker. Neither seems satisfactory. Both parties have given up something of value, and their sacrifices serve only to further complicate an already challenging evaluation of their dignity interests. It seems reasonable then, and necessary for maximizing social welfare, to consider the dignity interests

---

[103] Masterpiece Cakeshop, Ltd. v. Col. Civil Rights Comm'n, Docket No. 16-111 (U.S. argued Dec. 5, 2017).

[104] That the creation of the cake is a form of the baker's religious expression is a key assumption on which the case could turn. The forgoing discussion considers the consequences of its validity. If the court finds otherwise, it might be useful to imagine a case where a couple asks a religious painter to symbolically paint their union, or some other factual scenario where creative expression contrary to religious belief must be compelled.

[105] *See* Douglas Nejaime & Reva B. Segal, *Conscience Wars: Complicity-Based Conscience Claims in Religion and Politics*, 124 YALE L.J. 2516, 2579 (2015) (asserting that law directs lawmakers to consider "the harms to other citizens that accommodating complicity-based conscience claims may inflict"). For an argument that cases are rarely decided strictly on the costs and benefits to the parties, and instead involve an assessment of the costs and benefits to society at large, see Frank Fagan, *Renovating the Efficiency of Common Law Hypothesis*, THE TIMING OF LAWMAKING 280 (Frank Fagan & Saul Levmore eds. 2017) (developing a model where judges decide cases on the basis of allocative efficiency amongst parties and non-parties).

of non-parties or other societal norms and values.

In the *Piggie Park Enterprises* case, where the defendant pled denying service as a form of religious expression, the Supreme Court expressly invoked the advancement of the interests of non-parties in its decision to award attorney fees to the plaintiff.[106] Undoubtedly, the court based its decision on the text and purpose of civil rights legislation, but in doing so, it implicitly elevated a church-state separation norm above the norm of religious expression. The endorsement of public accommodation laws, when situated in similar conflicts, can also be understood as an affirmation of the predominance of a church-state separation norm. To the extent that this norm remains ascendant, it should be expected that religious expression will be removed from social interactions where dignity interests between parties are subjective, and where broader societal interests are implicated. Once the norm of free expression is reshaped, its conflict with civil rights will peter out, and it will begin its descent toward unconscious cognitive process unless it is called to conflict again by a norm entrepreneur.

Attitudes toward immigration, climate change, and religious expression subsist almost entirely on internalized values. Validity claims that only fuel interpersonal sanctioning operate on the surface of political team sports and make less of an impact on the shaping of preferences and the construction of deeply held beliefs. Regulating social media content to reduce those types of claims will do little to enhance social welfare. Aggregate levels of approval benefits and disapproval costs will remain relatively unchanged since the teams are engaged in zero-sum conflict. On the other hand, configuring network architecture with systemic social media regulation so that claims are funneled toward locations where discursive excellence thrives, can lead to greater levels of internalization. Consider that Reddit, in an effort to elevate the quality of its fora, "shadowbans" users who come to troll.[107] Traditional banning blocks the troll from the forum, but trolls can simply change their names and continue to troll. By contrast, shadowbanning blocks others from viewing the troll's messages: the troll continues to troll, sees her messages, and believes she is still trolling. But she is speaking to an empty hall, even as she believes

---

[106] Newman v. Piggie Park Enterprises, Inc., 390 U.S. 400, 402 (1968) ("If successful plaintiffs were routinely forced to bear their own attorneys' fees, few aggrieved parties would be in a position to advance the public interest by invoking the injunctive powers of the federal courts.").

[107] *See Can Someone Please Explain to Me What "Shadow Banning" Is?*, REDDIT, https://www.reddit.com/r/AskReddit/comments/11ggji/can_someone_please_explain_to_me_what_shadow/ (last visited Apr. 6, 2018).

the seats are full. Policing content rarely fosters rational deliberation and the internalization of facts and norms. By focusing on platform architecture, a tendency toward group polarization can be neutralized, and users can be nudged with systemic measures toward network locations where rational deliberation proliferates.

IV. THE REGULATION OF SOCIAL MEDIA

A. People and the First Amendment

The validity claims of people, lawmakers, and social media platforms are governed by different sets of rules even though their claims evolve and devolve in identical and predictable patterns. People, as norm entrepreneurs, make validity claims and engage in social enforcement. Claims, and enforcement of other people's claims, are speech acts governed by the First Amendment. Various types of speech—religious, political, commercial, obscene, and so on—receive various degrees of scrutiny, but all can be profitably analyzed with the general framework formulated by Learned Hand and later updated by Posner.[108] The essential idea is that law should compare the costs of forbidding speech with the costs of permitting it and choose the lesser of the two evils. By categorizing speech, law uses archetypes to identify the magnitude of costs in order to reach a decision. Forbidding religious or political speech, for instance, carries a greater cost than forbidding obscene speech. Conversely, permitting religious or political speech typically carries a lesser cost than permitting obscene speech. Religious and political speech can generate positive externalities by normalizing prosocial behavior and discourse; obscene speech can generate negative externalities by normalizing asocial behavior. The consideration of externalities introduces uncertainty: when speech is permitted, law cannot be sure that it will be socially costly. It must guess. Learned Hand's formula, therefore, directs law to forbid speech only if the costs of forbidding it are greater than the probabilistic costs of permitting it.[109]

Posner expands this idea by noting that the social costs of permitting speech may occur later in time.[110] Not only should social costs be discounted by their probability of occurrence, but they should additionally be discounted by the time that they take to arrive. If Martin Luther were governed by the First Amendment, he surely would have been permitted to post the *Ninety-Five Theses*. Not only did they constitute

---

[108] *See* US v. Dennis, 183 F.2d 201, 201 (5th Cir. 1950); Richard A. Posner, *Free Speech in an Economic Perspective*, 20 SUFFOLK U. L. REV. 2, 2 (1986).
[109] *See Dennis*, 183 F.2d at 201.
[110] *See* Posner, *supra* note 108 at 8.

high-value religious and political speech, but their disruptive effect (if considered a social cost) was hardly certain and took time.[111] With respect to the reconfiguration of religious expression set in motion by the *Theses*, any social cost would have certainly approached zero, given the length of time needed for its development. In general, speech acts that lead to the development of deep norms must be so heavily discounted for time and uncertainty that First Amendment law would counsel against their prohibition. People, as norm entrepreneurs engaged in shaping the future, are given free reign. This makes intuitive sense. Any decision in favor of defendants in *Masterpiece Cakeshop* will surely include a heavily-weighted rationale of the immediate social impact of permitting the baker to deny the couple service. This type of reasoning, while not strictly utilitarian, is expected from the "ideological" justices. What is interesting, is that the swing vote belongs to the justice who tends to engage in loose statements about time and tradition.[112] When considering probabilistic costs and benefits that occur later in time, a decision is simply more difficult and, unsurprisingly, swings.

Any assessment of First Amendment constraints would be incomplete without consideration of the costs of forbidding speech. These costs, which include the cost of suppressing valuable information plus any legal error generated from suppressing too little or too much, must be weighed against the social costs of permitting the information to be circulated. Returning to Luther, the costs of forbidding his post of the *Ninety-Five Theses* might have forestalled the development of a norm of religious expression and its evolution toward church-state separation. These costs, too, were sufficiently uncertain and would take many years

---

[111] *See id.*; *see also* Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) (permitting advocacy of unlawful conduct unless speaker intends to incite specific unlawful action that is likely to result imminently). Gregory notes that Luther initially called for an official council to reconcile differences and that his early goals were aimed at mild reforms within the confines of existing law and Roman Catholic authority. *See* GREGORY, *supra* note 102 at 57–58.

[112] For example, in *Romer v. Evans*, in reference to a Colorado law that fenced sexual minorities out of political processes, Justice Kennedy noted that: "It is not within our tradition to enact laws of this sort." 517 U.S. 620, 633 (1996). Comparing the approaches of Kennedy and Posner toward striking down prohibitions of gay marriage, Martha Nussbaum notes: "Posner's tone is skeptical, caustic, intolerant of cant; Kennedy's is solemn and lofty. Posner addresses concrete issues of welfare; Kennedy adduces high ethical abstractions, dignity and equality. Posner is punctilious in matters of argument, Kennedy loose and impressionistic." Martha C. Nussbaum, *Janus-Faced Law: A Philosophical Debate* in FRANK FAGAN & SAUL LEVMORE, THE TIMING OF LAWMAKING 270 (2017).

to incur. Quite obviously, Catholic authorities sought to suppress Protestant speech on the basis of its immediate effects, while reformers received protection from political leaders who benefitted immediately from agitation.[113]

B. Lawmakers and Politics

Lawmakers, as norm entrepreneurs, face an entirely different set of constraints to their capacity for making claims and enforcing the claims of others. Because claims and social sanctions are expressed in statutes, regulations, and judgments, lawmaker norm entrepreneurship is primarily demarcated by political feasibility. For some time, legal scholarship generally ignored the expressive power of law, which might be attributed to a comfortableness with its democratic pedigree: lawmakers, especially legislators, are accountable to the electorate. Besides, law's expressive power, when compared to law's immediate compliance-generating power, seems far less important: any expressive power can easily be short-circuited with new, countervailing rules.[114] The expressive function then, supports the underlying purpose of a rule by transmitting additional information beyond the rule itself. Because lawmaking involves opportunity costs, that is, lawmakers must choose to spend time and political capital on one rule versus another, grant certiorari, or develop particular regulations at the expense of others, lawmaking inherently conveys information about societal priorities. In addition, the public and participatory nature of lawmaking strengthens the salience of its object. Laws that are difficult to enforce because detection is costly express values which encourage victims to speak out; and laws that address social wrongs that affect a small number of people suggest a need for heightened public attention. In other cases, law might focalize a method for social coordination. But whatever the form it takes, legal expression is circumscribed by lawmakers' ability to pronounce law.

---

[113] *See* GREGORY, *supra* note 102 at 53–54 (noting that the protection of reformers was used as leverage in negotiations with Catholic authorities).

[114] Of course legal scholars did not entirely ignore the expressive function of law, but earlier articulations were more clearly formulated by sociologists. For instance, in his *History of Sexuality*, Foucault stated:

> I do not mean to say that the fades into the background or that the institutions of justice tend to disappear, but rather law operates more and more as a norm, and that the judicial institution is increasingly incorporated into a continuum of apparatuses (medical, administrative, and so on) whose functions are for the most part regulatory.

FOUCAULT, *supra* note 20 at 144.

C. Social Media Platforms and Regulation

The most apparent difference amongst people, lawmakers, and social media platforms is that people and lawmakers directly create content. Platforms are more like plumbers. They adjust their algorithms to control the flow of what their users see. By controlling the flow, social media platforms determine which validity claims are made over their networks and can exert meaningful control over patterns of approval, disapproval, pride, and guilt. For instance, Google Analytics closely tracked which issues its users fact-checked during the Obama-Romney presidential debates.[115] By tracking "who was searching what" during the debate, Google was able to deliver highly segmented advertising impressions of its users who had demonstrated, by means of their search histories, an elevated interest in a particular issue.[116] This is a clear example of a platform indirectly determining the contents of claims. Once these advertisements and other forms of claim-making are linked to engagement and amplification platforms such as Facebook and Twitter, users initiate patterns of social sanctioning. These patterns, too—of necessity—are controlled by the platforms. Social media users face opportunity costs of viewing and enforcing claims. They can only act on a limited number of messages. For this reason, social media platforms must limit and curate the messages that their users see. News stories, accompanying comments, and other forms of content are categorized and triaged. A user who checks the first few items of a news feed or Twitter account routinely ignores the items buried toward the bottom. Choice architecture and nudging are the natural outcomes of tailoring essentially limitless media to individual user profiles and characteristics.

The selection and prioritization of social media items, though algorithmically obscured, can be understood as motivated by profit and other managerial interests, which can sometimes present conflicts internal

---

[115] Google Analytics, Case Study, *Obama for America Uses Google Analytics to Democratize Rapid, Data-Driven Decision Making* (2013) https://static.googleusercontent.com/media/www.google.com/en//intl/hr_ALL/analytics/customers/pdfs/obama-2012.pdf.

[116] To the extent that Google Analytics was unable to identify which user was searching a particular issue, say, because a user was searching anonymously, then Google might be able to identify which issues were important to users from a particular geographic location by analyzing IP addresses. As noted in the case study, tailoring and directing messages on the basis of geography was critical for victory: "The results from Election Day speak for themselves: a resounding victory, with nearly every battleground state falling into the President's column. [Data analytics has been credited] for providing much of the winning margin." *Id.*

to the social media platform itself. These internal conflicts among managerial interests stand in contrast to the gubernatorial interests of the state.[117] To the extent that platforms engage in content moderation and censorship, they exercise quasi-judicial governance functions that can, in some cases, satisfy state preferences; however, this satisfaction is the outcome of alignment between platform and state interests.

For instance, when YouTube considers the removal of a terrorist recruitment video, it is considering a First Amendment question and exercising quasi-judicial power. It is obvious that managerial and gubernatorial interests can overlap. YouTube's profitability, corporate image, long-term viability, and capacity to avoid regulation depend on satisfactory operation of its quasi-judicial function. If YouTube fails to remove the video, its corporate image may decline, its users may go elsewhere, and the state may impose costly regulations. Inasmuch as managerial and gubernatorial interests strongly converge, there exists a good case for self-regulation.[118] Where they diverge, rules are desirable so long as their benefits exceed their costs.

This point is clearly seen in recent German legislation aimed at enforcing, on social media platforms used within Germany, speech content restrictions that were set in place there following the Second World War.[119] Embedded in Germany's federal criminal code are prohibitions against the "use of symbols of unconstitutional organizations" unless for "art or science, research or teaching;" speech that incites treason or other crimes; threats to commit various crimes; incitement to hatred, including through the dissemination of written materials; depictions of violence; and defamation of religions or religious and ideological organizations.[120]

Most social media platforms are globally sprawled and vast. Their

---

[117] *See* Balkin, *supra* note 13 at 1153 (2018) (noting that platforms engage in content moderation, which amounts to private governance); Klonick, *supra* note 13 at 1662 (referring to private content platforms as systems of governance).

[118] If platforms are unable or otherwise lack competence to carry out their overlapping interests, then their interests can be understood as insufficiently strong or weakly overlapping.

[119] Bundestag, *An Act to Improve Enforcement of the Law in Social Networks* (July 12, 2017) https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2. For the original version, see *Netzwerkdurchsetzungsgesetz*, 61 BUNDESGAZETZBLATT 3352–55 (2017) http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl117s3352.pdf.

[120] German Criminal Code § 86, 91, 111 100a, 111, 126, 130, 131, 140, 166, 184b.

managerial interests often fail to sufficiently overlap with localized gubernatorial interests such as Germany's.[121] Facing insufficient convergence, Germany cannot expect platforms to self-regulate in its interest. Moreover, Germany's benefit from the regulation is large because it enhances enforcement of an easily identifiable and constitutionally settled content restriction. Perhaps more importantly, the benefit is delivered with little cost to the state inasmuch as it relies on private citizens for its enforcement. Platforms with more than two million users in Germany must provide the public with "an easily accessible process" for registering complaints of illegal content directly with the platform.[122] For content that is "manifestly unlawful" the rule requires the platform to block access within twenty-four hours. For content that is simply "unlawful", the platform must block access within seven days unless "the unlawfulness of the content is dependent on the falsity of a factual allegation or is clearly dependent on other factual circumstances" or the platform "refers the decision regarding unlawfulness to a [lawfully] recognized self-regulation institution . . . and agrees to accept the decision of that institution."[123] If a platform receives more than 100 complaints per year, it must comply with various reporting requirements.[124] Users and members of the public who remain unsatisfied with the outcome of the complaint procedure, either because the platform is unresponsive, or they disagree with its decision, may file an online complaint with the German Department of Justice.[125] Without sufficiently strong convergence of managerial and gubernatorial interests, Germany chose to develop a rule that leverages average internet users for its enforcement and limits the general reporting requirements of platforms to aggregate and serious problems.

Note that this rule loosely tracks the *Restatement* approach to communication tort liability of distributors. Only if a platform is alerted of unlawful content and refuses to block it, can a user proceed to the German Department of Justice with the complaint.[126] By providing for notice-based

---

[121] *Cf.* Yahoo!, Inc. v. La Ligue Contre Le Racisme et L'Antisemitisme, 169 F. Supp. 1181, 1184 (2001) (noting that a French court found that Yahoo! violated § R645-1 of the French Criminal Code by permitting French users to purchase Nazi paraphernalia through its globally accessible website).

[122] Social Network Enforcement Act § 3.3.2

[123] *Id.* at § 3.2.3(a)–(b).

[124] *Id.* at § 2.

[125] *See* Bundesamt für Justiz [Federal Office of Justice], https://www.bundesjustizamt.de/DE/Themen/Buergerdienste/NetzDG/Service/Formulare/Formulare_node.html (last visited May 3, 2018).

[126] Compare RESTATEMENT (SECOND) OF TORTS § 581 (1) (1977) (attaching

liability, the rule avoids the complications presented by later decisions such as *Stratton Oakmont, Inc. v. Prodigy Servs. Co.*, which held distributors liable if they relied on automated editorial control.[127] Where it parts ways with American law, and in particular § 230 of the Communications Decency Act,[128] is that it prioritizes the gubernatorial interests of the state in content restrictions over the managerial interests of the platform such as content proliferation or cost minimization of aggressive enforcement.

On the other hand, scholars and courts have recognized that the immunity provision of § 230 encourages platforms to be good citizens, and remove offensive material, because they will not be held liable as editors if they do; and § 230 promotes free speech and e-commerce because platform immunity nurtures platform growth.[129] While some scholars have suggested that recognition of cyber civil rights should circumvent immunity, the regulatory benefit from speech suppression is clearer in Germany because specific content restrictions are codified.[130] In the United States, regulatory benefits must be developed by constitutional interpretation. Moreover, benefits can be interpreted to move in the opposite direction because the absence of regulation prioritizes a constitutionally guaranteed right of free speech. Only if the normative conflict is settled, will the gubernatorial interests of the state become clear.[131] For now, the state can avoid settling the conflict by allowing

---

liability to distributors if they know or should have known of the defamation).

[127] 1995 WL 323710 *5 (N.Y. Sup. Ct. 1995) (finding distributor liable as a publisher because it sought "to gain the benefits of editorial control" through automated content filtering and user guidelines for posting and that it "uniquely arrogated to itself the role of determining what is proper for its members to post and read on its bulletin boards.").

[128] 47 U.S.C. § 230 (c).

[129] *See* Klonick, *supra* note 13 at 1607 (cataloging court decisions and scholarship).

[130] On cyber civil rights, see Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61 (2009). Other rationales suggesting that the benefits of regulation are large include: cyberspace amplifies speech harms, especially sexual harassment; the anonymity of cyberspace has caused hate speech to go mainstream; and online harassment, bullying, and revenge pornography have proliferated. *See* Klonick, *supra* note 13 at *1614.

[131] This conflict has been present since the Founding. In argument against the Sedition Act of 1798, James Madison noted:

> [T]he unconstitutional power exercised over the press by the Sedition Act ought, more than any other, to produce universal alarm; because it is levelled against that right of freely examining public characters and measures, and of free communication among the people thereon, which

platforms to privately balance free speech values with other social goals such as robust security and the discouragement of hate speech. If the platforms do too much and thus infringe upon First Amendment principles, or too little and consequently degrade security and civil rights, the state can intervene. Where platform action converges with state preferences, a self-regulation model should prevail. It should be clear that platforms have some room to maneuver, so long as the gubernatorial interests of the state remain weakly defined. The state has room to maneuver inasmuch as it systemically regulates social media and avoids directly combatting speech acts.

The short- and medium-term profitability of platforms generally increases with their size and the presence of an unfettered legal environment. If users are unhappy with unmoderated content, they may go elsewhere. If the state is unhappy, it may impose rules. Indeed, much like lawmakers, platforms are accountable to the demands of their users and to broader political feasibility. Platforms accordingly intensify self-regulation up to the point where its marginal benefit equals its marginal cost. Setting an optimal content moderation policy maximizes the volume of content by balancing aggregate user engagement and alienation while avoiding costly state interference. Generally, if judges or legislators shoehorn social media platforms into paradigmatic company towns, television broadcasters, newspaper editors, municipal utilities, or governance institutions, it will likely be the result of a means-end instrumentalism for subsuming divergent managerial interests within broader societal concerns.[132]

As norm entrepreneurs, large platforms like Facebook, YouTube, and

---

> has ever been justly deemed the only effectual guardian of every other right.

James Madison, Report on the Virginia Resolutions (Jan. 1800) Writings 6:385-401, http://presspubs.uchicago.edu/founders/documents/amendI_speechs24.html.

[132] Company towns are functionally equivalent to state actors and must guarantee First Amendment rights. *See* Marsh v. Alabama, 326 U.S. 501, 502–03 (1945). Because radio and television broadcasters monopolize frequency spectrums, there exists a public right to suitable access and regulators are justified in requiring broadcasters to present both sides of public issues. *See* Red Lion Broad Co. v. FCC, 395 U.S. 367 (1969). Newspaper editors receive protection to decide the contents of their newspapers. *See* Miami Herald Pub. Co. v. Tornillo, 418 U.S. 241 (1974). On platforms as utilities, *see* Packingham v. North Carolina 137 S. Ct. 1730 (2017) (holding that preventing users from accessing social media platforms is a denial of a First Amendment right opening the door to treat platforms as quasi-utilities).

Twitter can develop longer-term projects delimited by prevailing social norms and political possibility. A freedom to pursue long-term managerial interests through norm entrepreneurship is consistent with First Amendment principles because an outcome must be deeply discounted by the uncertainty of its success and the time it takes to achieve it. Any justification for the limitation of platform construction of deep norms must be met with new law or creative lawyering. From a social welfare perspective, pursuit of deep norms remains unproblematic. Creating them, and eliciting habitual compliance, relies on the free choice of people to engage in internalization. In a world where the formation of political attitudes and beliefs are more like team sports, platforms cannot shape electoral outcomes insomuch as social facts and falsehoods matter little. On the other hand, platforms direct the flow of political messaging and can nudge users toward network locations where higher-quality political discourse is the norm.[133] If nudging decreases polarization costs, it may be worthwhile. While First Amendment doctrine circumscribes direct speech restrictions, the systemic regulation of platform architecture is more likely to survive constitutional scrutiny. In any case, platforms appear to be leading the way here in terms of developing creative architectures and implementing them, though one could imagine sustained divergence and the need for state action in the future.[134]

CONCLUSION

Social facts, norms, and falsehoods proliferate because of the actions of people and the architecture of platforms. Inasmuch as regulation is desirable, law should focus on systemic adjustment and reconfiguration of platform architecture and avoid targeting and suppressing speech contents. Rules that shape the contours of the forum, and the manner in which speech acts proliferate, can nudge speakers toward social media platform locations where discursive excellence thrives. Whether social media can be analogized to public utilities, company towns, broadcasters, newspaper editors, or governance institutions—and regulated accordingly—is important only to the extent that one of these models adequately aligns the managerial interests of the platform with the gubernatorial interests of the state. To the extent that these interests are already aligned, a self-regulation

---

[133] *See supra* note 99–100 and accompanying text.

[134] It should be clear that this Article focuses on content moderation and speech restrictions and sets aside for future work questions of private law. Inasmuch as platforms are violating end-user agreements or failing to take efficient precaution to safeguard user data, private contract and tort law claims may be enough to force platforms to internalize the external costs that these violations and breaches create. If not, then a case may be made for intervention.

model should prevail. Where they diverge, a case can be made for intervention.