Notes

Saving the Sinking Ship: How the United States Can Create an Effective Content Moderation Policy by Looking Abroad

ZHI YANG TAN[†]

Each day, the world creates another 2.5 quintillion bytes of data, with most of it being accessible by the average person through the smartphone they carry in their pocket. That data may often take the form of informative new articles or funny cat videos, but also hidden within that sea of information is content designed for more malicious purposes. While much of the world, and especially the U.S., has historically taken a laissez-faire approach to moderating online content, such an approach is quickly becoming outdated and ineffective as more people are exposed to disinformation or hate speech online, which can have effects that spill over into the real world. Governments and platforms are therefore facing the difficult problem of how to best limit this harmful content while not stifling the power of the internet as a tool for expression. Many other countries in the last decade have begun abandoning the laissez-faire approach and are developing their own solutions to online content moderation.

This Note presents an international typography of those approaches. It groups them into three general categories: platform-focused regulations meant to encourage platforms to properly moderate, user-focused regulations that punish citizens that create or disseminate harmful content, and education-based reforms that aim to create a more informed populace. Then, it examines in detail how each are implemented and their potential strengths and weaknesses. Finally, it proposes potential reforms for the U.S. that combines all three approaches in a way that empowers governments, platforms, and citizens themselves to address the problems cooperatively without engaging in state-sponsored censorship and abandoning important free speech principles in the process.

[†] J.D. Candidate 2022, University of California, Hastings College of the Law; Senior Communications Editor, *Hastings Law Journal*. The Author would like to thank Professor Chimène Keitner for her invaluable guidance and feedback throughout the entire writing process, as well as all the HLJ editors that reviewed this Note to make it the best it can be.

TABLE OF CONTENTS

INTRODUCTION

Since the beginning of the twenty-first century, the internet has become an integral part of everyday life for over half of the world's population.[1] It enables global communication, online marketplaces, and perhaps most important, content creation. In the last ten years, American internet usage has become increasingly mobile, with over 85% of Americans having access to the internet through the smartphone they carry around with them every day.[2] But this technological revolution comes with a cost. Those with ill intentions are given immense power to promulgate ideas that they otherwise would not be able to promulgate. This is because entry barriers that traditionally surrounded content creation suddenly came crumbling down, and now everyone has equal opportunity and power to create. As such, governments and private corporations around the world have been asking a big question: what is the best way to moderate user-created content online?

This Note addresses the various international approaches that have been taken towards solving the problem of content moderation problems, specifically regarding the rise of misinformation and disinformation online, and how those approaches could be applied to an American content moderation policy overhaul. Section I briefly discusses the roots of internet regulation in the U.S., and how a commitment to laissez-faire moderation has created an environment that encourages free expression but has also given rise to the use of the internet for malicious goals. Section II discusses the current landscape of the internet, and how its structure has allowed harmful content to thrive online. Section III discusses recent attempts to solve the content moderation problems, and various issues and considerations in content moderation that function as limitations on policy. Section IV analyzes the three main approaches to content moderation seen abroad and discusses the various benefits and drawbacks to each approach. And finally, Section V provides a framework for a possible solution in the U.S. centered around a combination of solutions enacted by governments and platforms alike.

I. THE HISTORY OF ONLINE CONTENT REGULATION

In the early days of the internet,[3] users, service providers, and regulators faced a problem: who was responsible for the content that was posted online? Legal precedent involving traditional print media held publishers liable for false or libelous content they published because they had direct control and

1. *Global Digital Population as of January 2021*, STATISTA (Sept. 10, 2021), https://www.statista.com/statistics/617136/digital-population-worldwide.

2. *Mobile Fact Sheet*, PEW RSCH. CTR. (Apr. 7, 2021), https://www.pewresearch.org/internet/fact-sheet/mobile.

3. While the Internet was originally developed as a military communications network, the Internet as we know it today, and for the purposes of this Note, truly began in 1989, when Tim Berners-Lee created the World Wide Web, allowing individual end-users to share information online. *History of the Web*, WORLD WIDE WEB FOUND., https://webfoundation.org/about/vision/history-of-the-web (last visited Jan. 24, 2022).

knowledge over the publications, while the distributors of those materials could not be liable because of their lack of control.[4] However, it was unclear how the roles of publisher and distributor would be applied on the internet, where websites often played both. In *Cubby, Inc. v. CompuServe, Inc.* ("*Cubby*"), the Southern District of New York held that CompuServe, an online information service that had "little or no editorial control" over the user-generated publications it hosted, was more akin to a distributor and was thus not liable for an allegedly libelous newsletter that it hosted on its servers.[5] However, in *Stratton Oakmont, Inc. v. Prodigy Services, Co.*, the New York Supreme Court in Nassau County held that Prodigy Services ("Prodigy"), another online information service, could be liable for a libelous post against Stratton Oakmont made by an unidentified bulletin board user because Prodigy was engaging in at least basic content moderation for the board.[6] Their use of content guidelines, software screening for offensive language, and the existence of an "emergency delete function" for when Prodigy felt that a post needed to be removed were enough for the court to differentiate it from *Cubby*, and hold that it was a publisher by virtue of its editorial control over the bulletin board content.[7]

These cases came to the attention of Chris Cox, a United States Representative from California, who thought that both cases were incorrectly decided.[8] He believed that forcing companies to choose between not moderating their content at all and avoiding any liability, or moderating their content to promote platform health and then being held liable for what users out of their direct control were posting, would turn the internet into the "Wild West[,] and nobody would have any incentive to keep the internet civil."[9] To solve this issue, Cox and fellow representative Ron Wyden from Oregon introduced the Communications Decency Act, which itself was an addition to the Telecommunications Act of 1996.[10] Section 230(c), of the act reads as follows:

> "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[11]

Often referred to as simply "Section 230," these words created an almost impervious shield for service providers against liability for content their users created.[12] Additionally, no service provider could be held liable for any good

4. Smith v. California, 361 U.S. 147, 153 (1959).

5. Cubby, Inc. v. CompuServe, Inc., 776 F.Supp. 135, 140 (S.D.N.Y. 1991).

6. Stratton Oakmont, Inc. v. Prodigy Servs., Co., 1995 WL 323710, at *2, *3 (N.Y. Sup. Ct. 1995).

7. *Id.* at *3–4.

8. Matt Reynolds, *The Strange Story of Section 230, the Obscure Law that Created our Flawed, Broken Internet*, WIRED (Mar. 24, 2019, 6:00 AM), https://www.wired.co.uk/article/section-230-communications-decency-act.

9. *Id.*

10. *Id.*

11. 42 U.S.C. § 230(c)(1).

12. *Section 230 of the Communications Decency Act*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/cda230 (last visited Jan. 24, 2022) ("In other words, online intermediaries that host or republish speech are

faith attempts to remove material that was obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable.[13] For the time being, this seemed to solve the dilemma—companies were now free to moderate user-generated content on their sites without the fear of being treated as the publisher of that content.

Section 230 greatly shaped the landscape of the early internet. By shielding companies from liability, it encouraged the development of internet intermediaries[14] with business models centered around providing access to an unlimited ocean of user-generated content, rather than their own novel content.[15] Because these protections were unique to American law at the time, and because places like Silicon Valley were becoming international hubs for the latest in technological innovation, the U.S. became the place to be for companies that wanted to create web-based services.[16] Some of the biggest intermediaries—Google, Facebook, Amazon—were founded and are still headquartered in the U.S., and therefore subject to U.S. laws.[17] In total, the U.S. economy's internet sector employs 425,000 people and contributes forty-four billion dollars to gross domestic product because of these and similar other protections.[18]

II. THE CURRENT LANDSCAPE OF THE INTERNET

But this explosive growth of the internet has its downsides. Before the internet, very few people had access to a platform with which they could broadcast their thoughts to the world. Now, anyone with an internet-connected device can access social media, post on forums, or write blogs about whatever is on their mind. As the Supreme Court held in *Reno v. American Civil Liberties Union*, the internet allows anyone to "become a town crier with a voice that resonates farther than it could from any soapbox."[19] And while this certainly was a powerful tool, and one that has been essential to the triumphs of democracies over dictatorships,[20] it has also enabled those with malicious intent to abuse that

protected against a range of laws that might otherwise be used to hold them legally responsible for what others say and do.").

13. 42 U.S.C. § 230(c)(2)(A).

14. Internet intermediaries are defined as companies that bring together or facilitate transactions between third parties on the internet. CHRISTIAN DIPPON, ECONOMIC VALUE OF INTERNET INTERMEDIARIES AND THE ROLE OF LIABILITY PROTECTIONS 3 (2017).

15. Stephen Engelberg, *Twenty-Six Words Created the Internet. What Will It Take to Save It?*, PROPUBLICA (Feb. 9, 2021, 2:00 PM), https://www.propublica.org/article/nsu-section-230.

16. *Section 230 of the Communications Decency Act*, *supra* note 12.

17. Sue Chang, *U.S. Companies Really Do Rule the Tech World—Here's the Chart to Prove It*, MARKETWATCH (Oct. 8, 2018, 12:07 PM), https://www.marketwatch.com/story/us-companies-really-do-rule-the-tech-worldheres-the-chart-to-prove-it-2018-10-08.

18. DIPPON, *supra* note 14, at 2.

19. Reno v. Am. Civ. Liberties Union, 521 U.S. 844, 870 (1997).

20. The internet and social media have been of particular importance in political revolutions in the last decade, including the Arab Spring. *See* Kali Robinson, *The Arab Spring at Ten Years: What's the Legacy of the Uprisings?*, COUNCIL ON FOREIGN RELS. (Dec. 3, 2020, 9:00 AM), https://www.cfr.org/article/arab-spring-ten-years-whats-legacy-uprisings.

power and create harmful content, and harm others on both the individual and community level.

Unlimited access to information has made it necessary for people to filter that information to easily digestible amounts. But that filtering process is never completely neutral, and users may trend towards creating their own personal echo chambers by self-selecting the information they want to see and ignoring the rest. Researchers from the Massachusetts Institute of Technology have found that shared partisanship (in their experiment, political ideology) has a causal effect on the social ties that people form with one another.[21] In another study, researchers found that Facebook users tended to "seek out and receive information that strengthens their preferred narrative."[22]

The effect of these echo chambers is worsened further through algorithms designed to keep users engaged with websites. Today, companies gather large sums of data on their consumers—their likes, dislikes, interests, and habits—that allows them to curate content that algorithms determine to be potentially interesting.[23] Progressive iterations will curate the feed further to ensure maximum engagement by the user, and by extension maximum profits for ad-supported platforms.[24] In the end, social media feeds become primarily populated with information that conforms to the user's biases,[25] further secluding them in the echo chamber.

And so, within these individualized echo chambers, an inherent desire to seek out information that conforms to one's beliefs and a system designed to feed a user with that information join forces to induce the spread of misinformation and disinformation, perhaps the most pervasive type of harmful content online today.[26] Trust in traditional news media has fallen significantly over the past few years, with only 46% of Americans in 2021 reporting that they trusted traditional media.[27] Instead, Americans are increasingly turning to social

21. Mohsen Mosleh, Cameron Martel, Dean Eckles, & David G. Rand, *Shared Partisanship Dramatically Increases Social Tie Formation in a Twitter Field Experiment*, 118 PNAS, no. 7, 2021, at 2.

22. Walter Quattrociocchi, Antonio Scala, & Cass Sunstein, Echo Chambers on Facebook 14 (June 13, 2016) (Discussion Paper No. 877) (on file with Harvard University, John M. Olin Center for Law, Economics, and Business), http://www.law.harvard.edu/programs/olin_center/papers/pdf/Sunstein_877.pdf.

23. Sang Ah Kim, *Social Media Algorithms: Why You See What You See*, 2 GEO. L. TECH. REV. 147, 147 (2017).

24. *Id.*; *see also* David Bauder & Michael Liedtke, *Whistleblower: Facebook Chose Profit Over Public Safety*, A.P. NEWS (Oct. 4, 2021), https://apnews.com/article/facebook-whistleblower-frances-haugen-4a3640440769d9a241c47670facac213.

25. Christopher Seneca, *How to Break Out of Your Social Media Echo Chamber*, WIRED (Sept. 17, 2020, 9:00 AM), https://www.wired.com/story/facebook-twitter-echo-chamber-confirmation-bias ("The algorithms ignore the recency and frequency of what our friends are posting and instead focus on what we 'like,' 'retweet,' and 'share' to keep feeding content that is similar to what we've indicated makes us comfortable.").

26. Misinformation is false or out-of-context information presented as fact. Disinformation is a type of misinformation created with the specific intent to deceive its audience. Meira Gebel, *Misinformation vs. Disinformation: What to Know About Each Form of False Information, and How to Spot them Online*, BUS. INSIDER (Jan. 15, 2021, 1:02 PM), https://www.businessinsider.com/misinformation-vs-disinformation.

27. Felix Salmon, *Media Trust Hits New Low*, AXIOS (Jan. 21, 2021), https://www.axios.com/media-trust-crisis-2bf0ec1c-00c0-4901-9069-e26b21c283a9.html.

media,[28] where the truth of posts can be hard to verify given their often user-generated nature. A 2019 survey by Ipsos found that 86% of people globally believe that they have been exposed to fake news, with another 86% of that group reporting initially believing what they saw.[29]

While a fake image of a former president soiling himself on a golf course[30] may be relatively harmless and perhaps even funny, the spread of misinformation and disinformation can have significant real-world effects when it encourages behavior that adversely affects others. This has been especially evident during the COVID-19 pandemic, where a wave of misinformation and disinformation surrounding the pandemic and various actions by health officials led international bodies to dub the phenomenon as its own "infodemic."[31] Almost two-thirds of Americans reported seeing news about the disease that seemed made up.[32] For example, a viral post from March 2020 alleging that a nationwide lockdown would be announced per the Robert T. Stafford Act[33] was one of many incidents that contributed to grocery store shelves being cleared out nationwide.[34] Studies have predicted that disinformation campaigns regarding the safety of COVID-19 vaccines may also be strong predictors of lowered vaccination rates in the future.[35]

More importantly, though, disinformation presents an existential threat to democracy. A system of government built on the will of the people requires citizens to be knowledgeable and informed so that they can make the best possible decisions. But this requirement can be and has been subject to abuse. For example, disinformation campaigns have been a favored tactic of Russia, which has used them since the fall of the Soviet Union.[36] While Russian disinformation has typically been targeted at its European interests, in the last

28. Darrell West, *How to Combat Fake News and Disinformation*, BROOKINGS. (Dec. 18, 2017), https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation.

29. Sean Simpson, *Fake News: A Global Epidemic Vast Majority (86%) of Online Global Citizens Have Been Exposed to It*, IPSOS (June 11, 2019), https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic.

30. Dan Evon, *Did President Trump Experience Diarrhea on a Golf Course?*, SNOPES (Apr. 10, 2017), https://www.snopes.com/fact-check/trump-diarrhea-golf-course.

31. World Health Organization, Novel Coronavirus (2019-nCoV) Situation Report – 13, at 2 (2020); António Guterres (@antonioguterres), TWITTER (Mar. 27, 2020, 8:55 PM), https://twitter.com/antonioguterres/status/1243748397019992065.

32. Christina Pazzanese, *Battling the 'Pandemic of Misinformation*,*'* HARV. GAZETTE (May 8, 2020), https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods.

33. Reuters Staff, *False Claim: A Text Says Trump to Declare Mandatory Quarantine Under the Stafford Act*, REUTERS (Mar. 18, 2020), https://www.reuters.com/article/uk-factcheck-quarantine-stafford-act/false-claim-a-text-says-trump-to-declare-mandatory-quarantine-under-the-stafford-act-idUSKBN2153UR.

34. James Peltz & Sam Dean, *Sales Are Up at Supermarkets. But that Brings New Problems for the Grocery Industry*, L.A. TIMES (Mar. 16, 2020), https://www.latimes.com/business/story/2020-03-15/grocery-store-industry-coronavirus.

35. Steven Reinberg, *Lies Spread on Social Media Hamper Vaccinations*, WEBMD (Oct. 30, 2020), https://www.webmd.com/vaccines/covid-19-vaccine/news/20201030/lies-spread-on-social-media-may-mean-fewer-vaccinations.

36. Neil MacFarquhar, *A Powerful Russian Weapon: The Spread of False Stories*, N.Y. TIMES (Aug. 28, 2016), https://www.nytimes.com/2016/08/29/world/europe/russia-sweden-disinformation.html.

three U.S. election cycles, Russia and other foreign countries have repeatedly attempted to create discord among U.S. voters through social media disinformation campaigns.[37] Notably, they did this by creating fake accounts to make posts that simply "stir the informational pot" by taking advantage of divisive topics.[38] While the effect these campaigns have had on political polarization is potentially limited,[39] they still highlight the fact that social media's ubiquity allows for attempts at large-scale democratic interference that can go virtually unnoticed on social media.

III. LOOKING FOR THE FIX

A. CONTEMPORARY SOLUTIONS TO CONTENT MODERATION

As of 2021, Section 230 is still controlling law. But it draws almost universal disdain from the political spectrum because it removes almost all liability for intermediaries. It notoriously shielded companies from liability when their services were used for threatening fellow employees,[40] housing discrimination,[41] and even coordinating terrorist attacks overseas.[42] The shield that was originally supposed to encourage moderation now seems to protect intermediaries even when they fail to take proper steps to prevent abuse of their services. Attempts at content moderation now seem to be prompted more so by societal outrage against the platforms, rather than a genuine desire to keep those platforms safe for users.[43]

The current state of online content moderation affairs is therefore a disjointed, laissez-faire system of internet governance, where the government has taken a back seat to let intermediaries themselves find a solution to moderate harmful content on their own platforms. Depending on the situation, platforms have been known to remove content, relocate content, directly edit content, add warnings for readers, add alternative perspectives, or disable comments.[44] But

37. NAT'L INTEL. COUNCIL, FOREIGN THREATS TO THE 2020 US FEDERAL ELECTIONS i (2021), https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf.

38. Joshua Yaffa, *Is Russian Meddling as Dangerous as We Think?*, NEW YORKER (Sept. 7, 2020), https://www.newyorker.com/magazine/2020/09/14/is-russian-meddling-as-dangerous-as-we-think.

39. Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, & Brendan Nyhan, *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*, HEWLETT FOUND. 15 (2018) ("Exposure to political disagreement on social media appears to be high, internet access and social media usage are not correlated with increases in polarization, and misinformation appears to have only limited effects on citizens' levels of political knowledge.") (citations omitted).

40. Delfino v. Agilent Techs., 145 Cal. App. 4th 790, 795 (Ct. App. 2006).

41. Chi. Laws' Comm. for C.R. Under Law v. Craigslist, Inc., 519 F.3d 666, 672 (7th Cir. 2008).

42. Force v. Facebook, Inc., 934 F.3d 53, 71 (2d. Cir. 2019).

43. Julia Carrie Wong & Olivia Solon, *Facebook Releases Content Moderation Guidelines – Rules Long Kept Secret*, THE GUARDIAN (Apr. 24, 2018, 6:14 PM), https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules ("The disclosure comes amid a publicity blitz by the company to regain users' trust following the Observer's revelation in March that the personal Facebook data of tens of millions of users was improperly obtained by a political consultancy.").

44. Eric Goldman, *Content Moderation Remedies*, MICH. TECH. L. REV. (forthcoming).

the application of these methods is often piecemeal and confusing, with vague definitions about what content is considered harmful and what should be done about it. Further complicating the issue is the fact that moderation is inherently influenced by the moderator's own beliefs and biases, which can introduce inconsistencies in enforcement, even within one platform.

One popular way for platforms to introduce consistency into their moderation processes is to publish universal community standards or guidelines: a set of rules or principles that dictate what kinds of content is considered appropriate. Every user is made aware of the community standards as part of their user agreement, and users agree to allow platforms to remove content they post that violates those standards.[45] But there is no standardized set of rules between platforms, leading to user confusion over what is appropriate or inappropriate on any given platform. For example, YouTube's policies contain an explicit ban on COVID-19 information that contradicts the WHO or any local health authorities, with videos found to violate this policy are *completely* removed from the site.[46] On the other hand, Facebook's community standards do not explicitly prevent the posting of misinformation, and instead simply note that posts that are internally deemed to contain fake news will be shown lower in user feeds.[47]

In addition, social media platforms have trended away from a model of meticulous community moderation to more large-scale commercial content moderation.[48] This was prompted by the sheer amount of material that needs to be screened for harmful content.[49] While sites like Reddit primarily use a combination of paid employees and volunteer moderators from the community, others like Twitter and YouTube now pride themselves on algorithmic moderation (separate from the algorithms that curate feeds) through systems that automatically classify content and determine whether they conform to community standards.[50] However, the workings of these systems are often black boxes that have no oversight or transparency by those outside of the company.[51]

---

45. *Terms of Service*, YOUTUBE, https://www.youtube.com/static?template=terms (last visited Jan. 24, 2022) ("If we reasonably believe that any Content is in breach of this Agreement or may cause harm to YouTube, our users, or third parties, we may remove or take down that Content in our discretion."); *Terms of Service*, TWITTER, https://twitter.com/en/tos (last visited Jan. 24, 2022) ("We reserve the right to remove Content that violates the User Agreement, including for example, copyright or trademark violations or other intellectual property misappropriation, impersonation, unlawful conduct, or harassment.").

46. *COVID-19 Medical Misinformation Policy*, YOUTUBE (May 20, 2020), https://support.google.com/youtube/answer/9891785.

47. *False News*, FACEBOOK, https://www.facebook.com/communitystandards/false_news (last visited Jan. 24, 2022).

48. Robert Gorwa, Reuben Binns, & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA & SOC'Y 2 (2020).

49. Katie Zigelman, *Why Use AI for Content Moderation*, SPECTRUM LABS (Aug. 10, 2019), https://www.spectrumlabsai.com/the-blog/2019/8/10/why-use-ai-for-moderation.

50. *Id.*

51. *Id.*

Notably, since 2016, various conservative figures have voiced concerns about moderation bias against their posts,[52] though whether that is truly the case is unclear.[53] And furthermore, algorithmic moderation is still in its infancy, leading to some of the same inconsistencies it was meant to solve.[54]

This, combined with a seeming lack of effectiveness in stopping the spread of harmful content, has led to criticism of this new trend towards algorithmic moderation. Tarleton Gillespie, Principal Researcher at Microsoft, argues that the shift towards these algorithms, especially with newer tech companies, has been partially motivated by venture capital.[55] Investors, especially those looking for the next tech unicorn, have a willingness to reward clever and novel software design, rather than systems that are effective at tackling the problems they were designed to solve.[56]

Platforms themselves are also voicing concerns that in the current regulatory environment, they are simply not well-suited to handle the sheer breadth and complexity of these moderation challenges on their own.[57] Facebook CEO Mark Zuckerberg has personally called for regulators to step in and guide companies like his own on the best way to engage in content moderation.[58] In the last few years, there have been various suggestions for fixing intermediary liability to encourage stronger content moderation, such as

52. Bill Chappell & Anastasia Tsioulcas, *YouTube, Apple and Facebook Ban Infowars, Which Decries 'Mega Purge'*, NPR (Aug. 6, 2018, 2:19 PM), https://www.npr.org/2018/08/06/636030043/youtube-apple-and-facebook-ban-infowars-which-decries-mega-purge; *see also* Prager Univ. v. Google LLC, 951 F.3d 991, 999 (9th Cir. 2020) ("PragerU prophesizes living under the tyranny of big-tech, possessing the power to censor any speech it does not like.").

53. Jennifer Graham, *Is Google Biased Against Conservatives?*, DESERET NEWS (Aug. 14, 2020, 10:00 PM), https://www.deseret.com/indepth/2020/8/14/21362500/is-google-biased-against-conservatives-breitbart-news-donald-trump-utah-mike-lee; Alison Durkee, *Are Social Media Companies Biased Against Conservatives? There's No Solid Evidence, Report Concludes*, FORBES (Feb. 1, 2021, 1:04 PM), https://www.forbes.com/sites/alisondurkee/2021/02/01/are-social-media-companies-biased-against-conservatives-theres-no-solid-evidence-report-concludes. Indiana Attorney General Todd Rokita announced on April 7, 2021 that the state would be launching an investigation into moderation practices. Lawrence Andrea, *AG Todd Rokita Investigating Big Tech Over What He Says is Conservative 'Censorship'*, INDYSTAR (Apr. 7, 2021, 7:55 AM), https://www.indystar.com/story/news/crime/2021/04/07/todd-rokita-investigating-facebook-google-apple-twitter/7120371002.

54. James Vincent, *YouTube Brings Back More Human Moderators After AI Systems Over-Censor*, THE VERGE (Sept. 21, 2020, 10:45 AM), https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns.

55. Sudhir Venkatesh, *"Someone Needs to Save the World from Silicon Valley"*, FREAKONOMICS: SUDHIR BREAKS THE INTERNET, at 25:50 (Apr. 26, 2020), https://freakonomics.com/podcast/someone-needs-to-save-the-world-from-silicon-valley.

56. *Id.*

57. Sebastian Herrera, *Tech Giants' New Appeal to Governments: Please Regulate Us*, WALL ST. J. (Jan. 27, 2020, 7:01 AM), https://www.wsj.com/articles/tech-giants-new-appeal-to-governments-please-regulate-us-11580126502 ("'These are the kinds of things that need to arrive at legislative solutions, versus individual CEOs of individual companies having to sort of come up with answers to what is a big, massive, societal challenge,' he said.").

58. Mark Zuckerberg, *The Internet Needs New Rules. Let's Start in These Four Areas.*, WASH. POST (Mar. 30, 2019), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.

assigning new duties to internet companies to their users,[59] but Congress has yet to agree on a course of action. Even since just the beginning of 2020, dozens of bills have been introduced in both houses, with each bill taking its own slightly different approach to the issue.[60] The 2020 bipartisan EARN IT (Eliminating Abusive and Rampant Neglect of Interactive Technologies) Act was possibly the most high-profile of these bills, but it received widespread public criticism for its potential to create a backdoor for law enforcement to access encrypted user conversations.[61]

The result of all of this? An environment where malicious users are often protected from the repercussions of their actions, and where governments are eager but powerless to hold the companies that give those users a platform legally responsible. This is undoubtedly the perfect environment for the creation and dissemination of harmful content. The ship that is the internet is fast sinking, but no one has the right tools or the individual power to save it. The 1990's approach of giving the intermediaries themselves protection and lenience has simply been outmoded by advances in technology and the growth of the internet, and thus a new solution is necessary.

B. IMPORTANT ISSUES FOR NEW REGULATIONS TO ADDRESS

Any new attempt to reform content moderation must first overcome a few key hurdles to implementation. First, and perhaps foremost, is a concern over authority: who should set the rules, and to what extent should platforms be regulated by third parties? While terms like misinformation or disinformation can have general definitions, there needs to be a consistent set of guidelines that platforms use to determine whether a specific piece of content meets that definition for there to be consistent enforcement across different platforms and different situations. Who will create those guidelines? Leaving it to a legislature or regulatory agency is a common approach, but it might not be the best solution. These bodies are often tasked with and strive to make rules that broad enough to allow for more flexible interpretation. This could result in a variety of outcomes, ranging from confusion over the meaning of certain standards, to potential abuse of vagueness to justify questionable moderation practices. Additionally, because people in these political bodies are distanced from the actual task of enforcement, and typically have no experience in the field themselves, it runs the risk of creating rules that look good on paper but are difficult to enforce. And lastly, the ideological makeup of these bodies changes as elections come and go,

59. MARK WARNER, POTENTIAL POLICY PROPOSAL FOR REGULATION OF SOCIAL MEDIA AND TECHNOLOGY FIRMS 6–7 (2018).

60. Kiran Jeevanjee, Brian Lim, Irene Ly, Matt Perault, Jenna Ruddock, Tim Schmeling, Niharika Vattikonda, & Joyce Zhu, *All the Ways Congress Wants to Change Section 230*, SLATE (Mar. 23, 2021, 5:45 AM), https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html.

61. Riana Pfefferkorn, *The EARN IT Act: How to Ban End-to-End Encryption Without Actually Banning It*, CTR. FOR INTERNET & SOC'Y AT STAN. L. SCH. (Jan. 30, 2020, 12:42 PM), https://cyberlaw.stanford.edu/blog/2020/01/earn-it-act-how-ban-end-end-encryption-without-actually-banning-it.

leading to potential instability in the rules with every change in administration. As such, requiring Congress, or perhaps the Federal Communications Commission, to develop detailed and binding standards for content moderation poses a significant amount of risk.

On the other hand, it may also not be wise to give the platforms themselves significant leeway in determining these guidelines. Currently, there are few limitations governing how Twitter, YouTube, or any intermediary chooses to establish its community guidelines. This gives platforms the freedom to tailor moderation to the specific platform and its users, but that too can cause issues. Corporations that are primarily driven by delivering profits to their shareholders and thus have an interest in gaining more power and control over their users through collected data[62] and may be influenced by ulterior motives separate and in contention with an interest in platform safety and integrity.[63]

The second main concern for potential solutions is the effect of intermediary liability laws and content moderation requirements on free speech. Justice Clarence Thomas argued in his *Biden v. Knight Institute* concurrence that "applying old doctrines to new digital platforms is rarely straightforward,"[64] and noted that private digital platforms wield a significant amount of power to cut off speech that could potentially be subject to First Amendment scrutiny.[65] The chief precedent in this area is *Marsh v. Alabama*, where the Supreme Court held that the First and Fourteenth Amendments could be applied to private actors as well as state actors if the private actor engages in the same actions that a government might.[66] Some argue[67] that social media platforms may fall under that rule, given that they play an increasingly important role in areas traditionally run by the government, like education[68] or elections.[69] In other cases, the Supreme Court has recognized cyberspace as a protected space under the First

62. Heather Kelly, *Google's Data Collection is Hard to Escape, Study Claims*, CNN BUS. (Aug. 21, 2018, 12:26 AM), https://money.cnn.com/2018/08/21/technology/google-data-collection/index.html.

63. Shannon Bond, *Over 400 Advertisers Hit Pause On Facebook, Threatening $70 Billion Juggernaut*, NPR (July 1, 2020, 11:44 AM), https://www.npr.org/2020/07/01/885853634/big-brands-abandon-facebook-threatening-to-derail-a-70b-advertising-juggernaut.

64. Biden v. Knight First Amendment Inst. at Colum. Univ., 141 S. Ct. 1220, 1221 (2021) (mem) (Thomas, J., concurring).

65. *Id.* at 1224–25.

66. Marsh v. Alabama, 326 U.S. 501, 506 (1946) ("The more an owner, for his advantage, opens up his property for use by the public in general, the more do his rights become circumscribed by the statutory and constitutional rights of those who use it.").

67. *See* Paul Domer, Note, *De Facto State Action: Social Media Networks and the First Amendment*, 95 NOTRE DAME L. REV. 893, 923 (2019).

68. Gerrit De Vynck & Mark Bergen, *Google Classroom Users Doubled as Quarantines Spread*, BLOOMBERG (Apr. 9, 2020, 4:45 AM), https://www.bloomberg.com/news/articles/2020-04-09/google-widens-lead-in-education-market-as-students-rush-online.

69. Sara Fischer, *Over 3 Million U.S. Voters have Already Registered on Social Media*, AXIOS (Sept. 21, 2020), https://www.axios.com/over-3-million-us-voters-already-registered-on-social-media-4db1b2eb-058e-43a9-899d-7f0ba8b49664.html.

Amendment because of how it enables communication for citizens and government officials alike.[70]

If these decisions are correct, and the First Amendment truly does limit the scope of moderation of speech made online, then platforms must be wary of how content moderation might infringe on a user's free speech rights. Pushing the limits of moderation too far may set off an avalanche of First Amendment challenges, especially if the government is playing too big of a role in said moderation. To be sure, traditional exceptions to the First Amendment like incitement or libel would also still apply in such a situation,[71] but apart from these particularized exceptions, moderation could be frustrated by an interest in protecting free speech.

IV. INTERNATIONAL APPROACHES TO CONTENT MODERATION

To find workable solutions, it may be helpful to look abroad for inspiration. Various foreign governments began taking steps towards playing a part in the content moderation process, leading to many different approaches and many different results. The United States can draw from these, and policymakers can use them to inform themselves about the costs and benefits of any course of action.[72] No foreign approaches are perfect, and many are deeply flawed, but they nevertheless provide case studies as to how different types of regulations might operate, and how public discourse and opinion could be affected.

The approaches can generally be grouped into three main categories: (1) regulation of platforms, (2) regulation of users, and (3) education to prevent the effects of harmful content. Importantly, the three are not mutually exclusive; countries often use a combination of some or all the approaches as part of a broader plan for regulation. However, this typology will focus on each named country's most prominent content moderation strategy.

A. PLATFORM-FOCUSED REGULATIONS

Increasing the amount of regulation for intermediaries is perhaps the most intuitive approach of the three. Much of the world originally took a cue from the early development of the internet in the U.S. and established liability shields like

70. *See* Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017); Knight First Amendment Inst. at Colum. Univ. v. Trump, 928 F.3d 226, 238 (2d. Cir. 2019), *vacated as moot*, Biden v. Knight First Amendment Inst. at Colum. Univ., 141 S.Ct. 1220 (2021).

71. *See* Knight, 928 F.3d at 237 (citing Brown v. Ent. Merchs. Ass'n, 564 U.S. 786, 790 (2011)) ("'[W]hatever the challenges of applying the Constitution to ever–advancing technology, "the basic principles of freedom of speech and the press, like the First Amendment's command, do not vary" when a new and different medium for communication appears.'").

72. Daphne Keller, *For Platform Regulation Congress Should Use a European Cheat Sheet*, THE HILL (Jan. 15, 2021, 1:00 PM) https://thehill.com/opinion/technology/534411-for-platform-regulation-congress-should-use-a-european-cheat-sheet.

Section 230.[73] But these days, if intermediaries and the services they provide are the conduits for the spread of harmful content, then they logically should be the focus of new regulation. This approach is by far the most popular internationally, and generally takes the form of legislation that imposes liability and new obligations on intermediaries and incentivizes them to engage in more moderation.

Most notably, the European Union recently proposed the Digital Services Act, designed to add new obligations to companies operating in Europe depending on their services and user base size.[74] The largest online platforms (defined as those that have over forty-five million users in Europe) would be subjected to the full gamut of obligations, including cooperation with national authorities, reporting criminal offenses that take place on their platforms, and greater transparency regarding how their content suggestion algorithms work.[75] Failure to comply with these new rules could lead to fines of up to six percent of the platform's annual revenue.[76]

Some national attempts, however, have been much harsher with regards to requirements placed on intermediaries, and highlight the potential risks of government overregulation. A primary example of this is the 2017 German "Netzwerkdurchsetzungsgesetz" or "NetzDG" law,[77] which was designed to keep internet companies responsible for the content shared through their services.[78] It did this by creating a litany of new duties for intermediaries. Companies would be required to establish new processes to deal with reports on their platforms regarding "manifestly unlawful content" (defined as a violation of one of twenty different provisions of the German Criminal Code), remove said content within twenty-four hours of receiving notice of a violating post, make monthly reviews of how reports are handled, and publish bi-annual reports on their processes.[79] A single failure to comply with any of these requirements could potentially result in a fifty million Euro fine.[80] However, the law drew criticism for encouraging companies to err on the side of caution and remove content more broadly rather than face the risk of a heavy fine.[81] United Nations

73. Liability shields were common in the early 2000s, but none went to quite the extent that Section 230 did. *See* Council Directive 2000/31/EC, art. 12, 2000 O.J. (L 178) 1, 12 (EC); Broadcastings Services Act 1992 (Cth) s 91(1) (Austl.); Electronic Communications and Transactions Act 25 of 2002 § 73-76 (S. Afr.).

74. Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM (2020) 825 final (Dec. 15, 2020).

75. *Id.* at Explanatory Memorandum.

76. *Id.* at art. 59.

77. Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], June 27, 2017 (Ger.).

78. While the law was originally aimed at social media companies, the definitions used are broad enough to capture many more internet platforms. *See Overview of the NetzDG Network Enforcement Law*, CTR. FOR DEMOCRACY & TECH. (July 17, 2017), https://cdt.org/insights/overview-of-the-netzdg-network-enforcement-law.

79. *Id.*

80. *Id.*

81. *Germany: Flawed Social Media Law*, HUM. RTS. WATCH (Feb. 14, 2018), https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

Special Rapporteur for the Protection of Freedom of Opinion and Expression ("Special Rapporteur") David Kaye also claimed that the law's restrictions on speech without a system of judicial oversight was not compatible with international human rights law.[82]

NetzDG's effect has not been limited to just Germany and German content; it has also been described by think tank Justitia as a "prototype" for similar platform moderation laws around the world.[83] In the years that followed NetzDG's passage, over a dozen different countries passed broad laws removing intermediary immunity and imposing content moderation obligations on platforms.[84] Almost all of them cited or referenced NetzDG as inspiration.[85] Some of these laws, like those in Vietnam and Russia, even contain language that directly mirrors specific clauses found in NetzDG.[86] The most notable of these was the May 2020 French "Loi Avia," which similarly required platforms to take down "manifestly illegal content" within twenty-four hours of a report and imposed harsh fines on platforms that failed to do so.[87] However, the law was later struck down in almost its entirety by the French Constitutional Council.[88]

Instead of directly doling out these harsh punishments onto platforms, what about a new regulatory body to help guide and oversee them? Some have considered this approach, but public response has been mixed. For example, in 2019, the United Kingdom, also inspired by NetzDG,[89] suggested in its "Online Harms White Paper" new laws that would create an independent regulatory body (funded by a tax on internet platforms) that would have the legal authority to establish standards for what is and is not permissible online, and the authority to take action against platforms that fail to comply.[90] But this approach has also been criticized as state regulation of speech[91] and creating financial and legal

82. David Kaye (Special Rapporteur for the Protection of Freedom of Opinion and Expression), *Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, OL DEU 1/2017 (June 1, 2017).

83. JACOB MCHANGAMA & JOELLE FISS, THE DIGITAL BERLIN WALL: HOW GERMANY (ACCIDENTALLY) CREATED A PROTOTYPE FOR GLOBAL ONLINE CENSORSHIP 6 (2019).

84. *Id.*

85. *Id.* at 6–16.

86. *Id.* at 8, 13.

87. Jacob Schulz, *What's Going on with France's Online Hate Speech Law?*, LAWFARE (June 23, 2020), https://www.lawfareblog.com/whats-going-frances-online-hate-speech-law#.

88. Aurelian Breeden, *French Court Strikes Down Most of Online Hate Speech Law*, N.Y. TIMES (June 18, 2020), https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html.

89. MCHANGAMA & FISS, *supra* note 83 at 12.

90. JEREMY WRIGHT & SAJID JAVID, ONLINE HARMS WHITE PAPER 53–63 (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973939/Online_Harms_White_Paper_V2.pdf. However, in 2020, the UK government announced that it would instead bestow these powers upon the country's Office of Communication. *Regulator Ofcom to Have More Powers Over UK Social Media*, BBC (Feb. 12, 2020), https://www.bbc.com/news/technology-51446665.

91. Alex Hern, *Internet Crackdown Raises Fears for Free Speech in Britain*, THE GUARDIAN (Apr. 8, 2019), https://www.theguardian.com/technology/2019/apr/08/online-laws-threaten-freedom-of-speech-of-millions-of-britons.

barriers for smaller companies that might have difficulty implementing new compliant systems.[92]

The scope of platform-focused regulation is an important sticking point for most of these approaches, and governments must be careful in their attempts to not overregulate. But perhaps the more important issue with this approach is something even more fundamental: the fact that they are designed to make moderation and liability platform-centric. Economist Steven Levitt remarked in an interview with YouTube CEO Susan Wojcicki that devoting resources to implement systems that simply *remove* harmful content from platforms does not properly account for the fact that content moderation is, in the language of game theory, a "repeated game."[93] A repeated game is one where there are many, sometimes infinite, instances of the same interaction over a period of time;[94] exactly the kind of situation that platforms and regulators face when there is a constant stream of new content being posted. Consistently playing the game of moderation with offenders and developing countermeasures may work in the short run to remove content soon after posting, but in the long run, a proper solution would need to end the game entirely. It needs to cut off the flow of harmful content onto platforms, rather than trying to clean up spills as fast as possible.

B. USER-FOCUSED REGULATIONS

That may be why some countries have instead, as Levitt posited would be a much more efficient use of resources, opted to focus on laws that punish individual users or organizations for creating harmful content. By creating this disincentive, these countries seek to stop the problem at its source and deter users from generating the offending content in the first place. Theoretically, doing so would thereby remove the bad actors, leading to a lessened need for platforms and policymakers to engage in moderation in the first place and keeping discussion online more open and unrestricted.

This user-focused approach has seen a growth in popularity throughout the last few years. In 2019, Russia passed a law holding individuals that spread what the government considers to be false information liable for up to approximately $7,600 in fines, with even stricter fines if the information causes injury, death, or a disturbance of public order.[95] In 2020, seemingly in response to a rise in COVID-19 misinformation, those penalties were made even more severe, with

92. Adam Satariano, *Britain to Create Regulator for Internet Content*, N.Y. TIMES (Feb. 12, 2020), https://www.nytimes.com/2020/02/12/technology/britain-internet-regulator.html.

93. Steven D. Levitt, *Susan Wojcicki: "Hey, Let's Go Buy YouTube!"*, FREAKONOMICS, at 23:57 (Oct. 16, 2020), https://freakonomics.com/podcast/pima-susan-wojcicki.

94. *Game Theory III: Repeated Games*, POLICONOMICS, https://policonomics.com/lp-game-theory3-repeated-game (last visited Jan. 24, 2022).

95. Astghik Grigoryan, *Russia: Russian President Signs Anti-fake News Laws*, LIB. OF CONG. (Apr. 11, 2019), https://www.loc.gov/item/global-legal-monitor/2019-04-11/russia-russian-president-signs-anti-fake-news-laws.

individuals facing up to approximately $25,000 in fines and up to five years in prison.[96] Malaysia passed a similar law in 2018 that would impose up to approximately $119,000 in fines and prison terms of up to six years.[97] While that specific legislation was repealed soon after the party in control of parliament changed,[98] the new government then turned around and used emergency powers granted for COVID-19 pandemic purposes to enact a similar ordinance only a few years later.[99] Neighboring Singapore passed its own version of the law in 2019, allowing *any* government minister to unilaterally declare a statement to be false and issue authorities to take action.[100] It also added the ability for Singaporean officials to extend their reach extraterritorially, as long as the communication itself is made "in Singapore."[101]

China is especially notorious for cracking down on dissidents, and its actions and approach in the context of harmful content online are no different. For over two decades, the Measures for Security, Protection, and Administration of the International Networking of Computer Information Networks have outlawed the creation and spread of harmful content online, including "information that fabricates or distorts facts, spreads rumors and disrupts social order," and "information that openly insults others or fabricates facts to slander others."[102] These regulations were later re-codified under its 2016 Cybersecurity Law.[103] But despite its strict regulation of what citizens can and cannot read or write, it has seen a significant rise in "rumors," the word used to describe what Americans might call fake news.[104] To combat this, it has required users to register their real identities with internet service providers,[105] and has amended its criminal laws to punish the spread of rumors with up to a seven-year prison

96. Daria Litvinova, *Russia Fines Opposition Radio Station for Fake News*, A.P. NEWS (June 19, 2020), https://apnews.com/article/47b0ee05dd531c693c860e4c05766775.

97. Anti-Fake News Act, Act 803, pt. II, s 4(1), (Apr. 9, 2018) (Malay.).

98. Reuters Staff, *Malaysia Parliament Scraps Law Penalizing Fake News*, REUTERS (Oct. 9, 2019), https://www.reuters.com/article/us-malaysia-politics-fakenews/malaysia-parliament-scraps-law-penalizing-%20fake-news-idUSKBN1WO1H6.

99. Joseph Sipalan, *Malaysia Defends Coronavirus Fake News Law amid Outcry*, REUTERS (Mar. 12, 2021, 12:06 PM), https://www.reuters.com/article/malaysia-politics/malaysia-defends-coronavirus-fake-news-law-amid-outcry-idUSL4N2LA2EX.

100. Protection from Online Falsehoods and Manipulation Act, No. 18 of 2019, pt. 3 s 10 (Sing.).

101. *Id.* at pt. 2, s 7 (Sing.).

102. Jisuanji Xinxi Wangluo Guoji Lianwang Anquan Baohu Guanli Banfa (计算机信息网络国际联网安全保护管理办法) [Measures for Security, Protection, and Administration of the International Networking of Computer Information Networks] (promulgated by the Ministry of Public Security, Dec. 30, 1997, effective Dec. 30, 1997), at art. 5, *translated in* LAWINFOCHINA, https://www.lawinfochina.com/display.aspx?id=6247&lib=law.

103. Zhonghua Renmin Gongheguo Wangluo Anquan Fa (中华人民共和国网络安全法) [Cybersecurity Law of the People's Republic of China] (promulgated by the Standing Committee of the National People's Congress, Nov. 7, 2016, effective June 1, 2017), art. 12, *translated in* LAWINFOCHINA, https://www.lawinfochina.com/display.aspx?id=22826&lib=law#.

104. LANEY ZHANG, GRACELA RODRIGUEZ-FERRAND, EDOUARDI SOARES, TARIQ AHMAD, & LANEY ZHANG, LAW LIBRARY OF CONGRESS, INITIATIVES TO COUNTER FAKE NEWS IN SELECTED COUNTRIES 18 (2019), https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1181&context=scholcom.

105. *Id.* at 20.

sentence.[106] Network operators also must keep track of users and their posts, and report to the authorities if a violation is detected.[107] And recently, citizens were encouraged to "enthusiastically report harmful information" by other citizens who "spread 'mistaken opinions.'"[108]

While these four countries are not the only ones to implement user-focused moderation laws, all four do share one commonality that highlights an inherent danger in taking this approach: they all have governments notorious for engaging in the suppression of divisive or unpopular speech. Freedom House's 2020 indices rated these four countries as either not free (Russia, China) or only partly free (Malaysia, Singapore) in Global Freedom and Internet Freedom.[109] While strict punishments for the spread of harmful content can act as a powerful deterrent, they require governments to restrain themselves from broad application. Many of these laws define their terminology extremely vaguely, leaving citizens unsure of the exact boundaries of acceptability on the internet. Therefore, there exists a fear that other authoritarian governments around the world will use digital concerns to justify similarly vague laws, affording them even greater control over what their citizens can access online regardless of whether that content is truly harmful.[110] And like with platform-focused regulations, when the rules are not clear to users, it creates a chilling effect on speech.[111] Many may choose to err on the side of caution and restrain themselves from speaking about topics that could potentially get them in trouble.

User-focused laws also may implicate potential human rights violations while trying to prevent the spread of harmful content. The United Nations has taken a leading role in espousing this view, insisting that government policies online must be crafted in a way that respects the rights of individuals.[112] For example, it noted in 2018 that legislation implemented by a member state, facially designed to criminalize the sharing of false information online, was instead used to silence dissent, and thus "cast[ed] a hostile shadow over the exercise of civil liberties."[113] The Special Rapporteur also recommended that "[s]tates should repeal *any* law that criminalizes or unduly restricts expression,

106. *Id.* at 19.

107. *Id.* at 20–21.

108. Lauren Giella, *China Encourages Citizens to Report Each Other for Posting 'Mistaken Opinions' on Internet*, NEWSWEEK (Apr. 19, 2021, 11:07 AM), https://www.newsweek.com/china-encourages-citizens-report-each-other-posting-mistaken-opinions-internet-1584696.

109. *Freedom on the Net*, FREEDOM HOUSE, https://freedomhouse.org/report/freedom-net (last visited Jan. 24, 2022) (scroll down; then click "Explore the Map"; then click on either "Internet Freedom" or "Global Freedom").

110. Adrian Shahbaz, *Freedom on the Net 2018: The Rise of Digital Authoritarianism*, FREEDOM HOUSE, https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism (last visited Jan. 24, 2022).

111. *Reno*, 521 U.S. at 871–72 ("[T]he CDA is a content-based regulation of speech. The vagueness of such a regulation raises special First Amendment concerns because of its obvious chilling effect on free speech.").

112. Wafa Ben-Hassine, *Government Policy for the Internet Must be Rights-Based and User-Centered*, U. N. CHRON., https://www.un.org/en/chronicle/article/government-policy-internet-must-be-rights-based-and-user-centred (last visited Jan. 24, 2022).

113. *Id.*

online and offline," in the interest of protecting platforms of public expression.[114]

C. EDUCATION-BASED SOLUTIONS

One possible way to get around the issues that come with directly punishing users who post harmful content is to instead focus on limiting the harm they can cause to the public. The goal with this approach seems to be to create an enlightened version of the "marketplace of ideas,"[115] where not only the ideas themselves clash to leave truth alone standing, but where the individual also has the competency to parse all the information and come to their own informed conclusions. To accomplish that, governments generally engage in public initiatives that raise awareness about certain topics or educate individual users on how to avoid the dangers of harmful content on the internet.

Nordic countries have led the way with these types of initiatives by establishing cooperative efforts throughout different sectors of society.[116] For example, Sweden has engaged in a massive effort to educate its citizens about misinformation, rather than trying to stop its spread with legislation.[117] Its Civil Contingencies Agency envisioned as early as 2013 a potential future in which "it may be difficult to distinguish public relations from news and [where] rogue news sources could have a major impact."[118] As part of its efforts to prevent this scenario, it published a 2018 pamphlet detailing how to counter "information influence activities."[119] The pamphlet details how communicators, such as public officials and influential organizations, can recognize and combat these activities.[120] It teaches what to look for in news articles, how to recognize bots, and how to address a target audience to counter the misinformation or disinformation.[121] Another pamphlet was created for the general population, urging citizens to be on the lookout for false information and to critically

114. David Kaye (Special Rapporteur), *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) (emphasis added).

115. The "marketplace of ideas" was a theory by John Stuart Mill espousing that because no one can know the truth, ideas in a free society could compete in the marketplace for truth and acceptability. Better ideas would be more successful in this marketplace, while bad, harmful, or outdated ideas would naturally fail. David Schultz & David Hudson, *Marketplace of Ideas*, FREE SPEECH CTR. (June 2017), https://www.mtsu.edu/first-amendment/article/999/marketplace-of-ideas.

116. MARTINA CHAPMAN, MAPPING OF MEDIA LITERACY PRACTICES AND ACTIONS IN EU-28 at 134–44, 170–81, 338–47 (Maja Cappello, Francisco Javier Cabrera Blázquez & Sophie Valais eds., 2016), https://rm.coe.int/1680783500.

117. OLGA ROBINSON, ALISTAIR COLEMAN, & SHAYAN SARFARIZADEH, A REPORT OF ANTI-DISINFORMATION INITIATIVES 4 (2019), https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/08/A-Report-of-Anti-Disinformation-Initiatives.pdf.

118. SWEDISH CIV. CONTINGENCIES AGENCY, FIVE CHALLENGING FUTURE SCENARIOS FOR SOCIETAL SECURITY 25 (2013), https://www.msb.se/siteassets/dokument/publikationer/english-publications/five-challenging-future-scenarios-for-societal-security.pdf.

119. SWEDISH CIV. CONTINGENCIES AGENCY, COUNTERING INFORMATION INFLUENCE ACTIVITIES 7 (2018), https://www.msb.se/RibData/Filer/pdf/28698.pdf.

120. *Id.*

121. SWEDISH CIV. CONTINGENCIES AGENCY, *supra* note 118, at 22, 24, 32.

appraise all sources.[122] And finally, these educational efforts have also been aimed at the country's youngest demographic, with popular cartoon character Bamse the Bear being recruited to help teach young children about the dangers of fake news.[123]

Other countries around the world have also pursued education to combat disinformation. Kenya, in partnership with the U.S. Embassy in the country, established the Young African Leaders Initiative (YALI) Checks campaign in 2018 to help young Kenyans learn to spot misinformation.[124] It teaches a three-step process: (1) stop before you share, (2) reflect on what you see or read, and (3) verify that the information is accurate.[125] The campaign website provides quizzes, workbooks, and games to help users understand and improve on their media literacy skills, and various events over the course of the program's first year provided additional support and learning opportunities.

So far, the educational approach seems to be quite effective. A few years after their programs went into effect, Nordic countries have ranked among the highest in Europe for media literacy,[126] and young Kenyans were ranked in the upper midrange for African countries.[127] But while the results are clear, the reasons for this success are not. For example, Finland (another Nordic country that has been seen internationally as a paragon for how to fight fake news) ranks highest in many other quality-of-life indices as well, causing some to interpret this fostering a better environment for education approaches, thus leaving less room for malicious actors to use misinformation to sow division.[128] Yet that fails to explain how Kenya, with a quality-of-life that is relatively much poorer[129] still managed to rank highly in comparison to other African countries. Further research should be conducted on the efficacy of these programs in different regions around the world and what factors are contributing to success.

---

122. SWEDISH CIV. CONTINGENCIES AGENCY, IF CRISIS OR WAR COMES 6 (2018), https://www.dinsakerhet.se/siteassets/dinsakerhet.se/broschyren-om-krisen-eller-kriget-kommer/om-krisen-eller-kriget-kommer---engelska-2.pdf.

123. Lee Roden, *Why This Swedish Comic Hero Is Going to Teach Kids About Fake News*, THE LOCAL (Jan. 16, 2017), https://www.thelocal.se/20170116/why-this-swedish-comic-hero-is-going-to-teach-kids-about-fake-news-bamse.

124. *Ambassador Godec and U.S. Embassy Counter Fake News with Media Literacy Campaign*, U.S. EMBASSY IN KENYA (Mar. 14, 2018), https://web.archive.org/web/20190319142019/https:/ke.usembassy.gov/ambassador-godec-u-s-embassy-counter-fake-news-media-literacy-campaign/.

125. *YALIChecks*, YOUNG AFR. LEADERS INITIATIVE, https://yali.state.gov/checks (last visited Jan. 24, 2022).

126. MARTIN LESSENKI, JUST THINK ABOUT IT. FINDINGS OF THE MEDIA LITERACY INDEX 2019, at 5 (2019).

127. *Critical but Less Creative: Media and Information Literacy Amongst Kenya's Youth*, DW AKADEMIE (Oct. 22, 2020), https://www.dw.com/en/critical-but-less-creative-media-and-information-literacy-amongst-kenyas-youth/a-55273817.

128. Eliza Mackintosh, *Finland Is Winning the War on Fake News. What It's Learned May Be Crucial to Western Democracy*, CNN (May 2019), https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/.

129. Kenya is ranked 143 on the UN Human Development Index, compared to Finland at rank 11. *Latest Human Development Index Ranking*, U.N. DEV. PROGRAMME, http://hdr.undp.org/en/content/latest-human-development-index-ranking (last visited Jan. 24, 2022).

V. APPLYING THESE APPROACHES TO THE UNITED STATES

A. GOVERNMENT-ENACTED SOLUTIONS

The current content moderation system in the United States is not working. Citizens, policymakers, and even the internet intermediaries themselves all agree on this one point. As such, it is time for the government to step in and assist with moderation efforts. Foundationally, this may be at odds with the idea of the United States' prized free market system, where companies should be at liberty to do what they want on their platforms. But the reality is that allowing the system to remain relatively free from government regulation has not resulted in a market solution to the dissemination of harmful content in the more than two decades since consumer internet access has become ubiquitous. Platforms have been pressured for years to find and implement solutions on their own, but the recent spike in disinformation shows that we are still far away from the day that such an approach can work.

Even the optimistic scenario in which platforms can independently moderate their content with reasonable success may not be a desirable one. The largest tech companies today possess near-monopolies on some of the most commonly used channels of communication and information dissemination.[130] For example: this Note was inspired by posts made by a government official on Twitter;[131] it was written and edited using software made by Microsoft;[132] and research on the competing views and approaches in this area was done using Google's search engine[133] on sites with content that is likely hosted on either Amazon, Google, or Microsoft's cloud services.[134] Allowing these massive private companies to unilaterally decide, with minimal requirements for accountability or transparency, what is and is not allowed on the internet, may lead to the rules of the web being determined by market forces and business decisions instead of morality and ethics.[135]

---

130. Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L. REV. 1353, 1385 (2018); Brian Fung, *Congress' Big Tech Investigation Finds Companies Wield 'Monopoly Power'*, CNN BUSINESS (Oct. 6, 2020, 9:12 PM), https://www.cnn.com/2020/10/06/tech/congress-big-tech-antitrust-report/index.html.

131. Ajit Pai (@AjitPai), TWITTER (Oct. 15, 2020, 11:30 AM), https://twitter.com/ajitpai/status/1316808733805236226?lang=en.

132. Microsoft Windows possesses 88% of the personal computer operating system market share. Nat Levy, *Windows 10 Market Share Passes 50% as Microsoft Continues to Dominate Traditional PC Market*, GEEKWIRE (Sept. 3, 2019, 6:53 AM), https://www.geekwire.com/2019/windows-10-market-share-passes-50-microsoft-continues-dominate-traditional-pc-market ("The latest figures show how dominant Windows remains in the PC arena, with a market share of roughly 88 percent.").

133. Google's search engine accounts for more than 90% of all web searches made. Jeff Desjardins, *How Google Retains More Than 90% of Market Share*, BUS. INSIDER (Apr. 23, 2018, 4:35 PM), https://www.businessinsider.com/how-google-retains-more-than-90-of-market-share-2018-4.

134. Felix Richter, *Amazon Leads $150-Billion Cloud Market*, STATISTA (July 5, 2021), https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers.

135. Langvardt, *supra* note 130, at 1387 ("Under such a system, the shape of free speech will be determined by popular opinion, market pressures, governmental pressures, and managerial conscience."); Rob Reich, Mehran Sahami & Jeremy M. Weinstein, *Facebook Isn't the Only Problem*, CNN BUS. (Oct. 14, 2021),

But while some government regulations will be necessary, regulations as strict as those seen abroad are likely not compatible with American society and its values. A NetzDG-esque system of strict rules and heavy fines focused on punishing platforms could stunt the growth of online communications, and even potentially run afoul of the Constitution as either a limitation on free speech or an excessive fine. Alternatively, attempts to crack down on dangerous users through lengthy prison terms and fines could go beyond necessary moderation for safety and instead become state suppression of speech. New regulation certainly needs to be enacted, but it must be done so in a way that minimally impacts the country's longstanding principles regarding speech and business operation. We can do this through a three-pronged approach involving a restructuring of the government's platform-based approach, a moderate increase in the user-focused disincentives for disinformation, and a heavy cross-sector push for more media literacy education.

1. Amending Section 230

First, platform-focused regulations must undoubtedly resolve the Section 230 issue. While the foundational principles of the law are sound, it gives intermediaries too much protection for too little work and acts as a barrier to keeping them responsible for the things that happen on their platforms. During his presidential campaign, President Biden called for Section 230 to be revoked, but only for companies that knowingly propagate falsehoods.[136] This likely would not work for two reasons. First, the standard President Biden proposed seems simple on its face but would be confusing in practice. In the era of automated content algorithms that push out exactly the content that users are seeking themselves, what does it even mean to "knowingly propagate" a falsehood? One might say that Facebook or Twitter are not necessarily actively *choosing* to spread disinformation, as opposed to users abusing their platform's automated technology to do so (although President Biden argued that Facebook was knowingly propagating falsehoods).[137] And if a platform can only be held liable if, say, a human makes the final decision, what happens to a platform when its human content moderator makes a close judgment call, or its software engineer tweaks the algorithm? Basing the standard on subjective knowledge could create many difficult situations and unintended consequences.

Second, the standard seems to conflict with the good faith exception and may disincentivize moderation. There is an immense amount of content that is

https://www.cnn.com/2021/10/14/perspectives/facebook-frances-haugen-big-tech-regulation/index.html ("But the push to scale new technologies quickly and achieve market dominance makes it even more likely that societal harms aren't fully considered until the negative consequences become evident and inescapable.").

136. Editorial Board, *Joe Biden Former Vice President of the United States*, N.Y. TIMES (Jan. 17, 2020), https://www.nytimes.com/interactive/2020/01/17/opinion/joe-biden-nytimes-interview.html.

137. *Id.* ("It should be revoked because [Facebook] is not merely an internet company. It is propagating falsehoods they know to be false . . . .").

posted daily on the internet.[138] Why would a platform bother to moderate any of it in good faith and work hard to protect its users when that good faith could just as easily be labeled as propagation? Platforms would then run the risk of making a mistake (or even simply an unpopular decision), being stigmatized for it, and losing all their Section 230 protection. The rational-actor CEO of any platform in such a situation would instead choose to turn a blind eye to content moderation beyond what is necessary, bringing the whole issue back to where it started in 1996.

Section 230 should stand but be amended to create a higher bar than "good faith" necessary to obtain the exemption for moderation.[139] As is, the exemption is inherently subjective and allows for an "I'm trying my best" defense to liability for platforms to fall back on. We live in the age of information and data, and the standards for liability should reflect this. Therefore, Congress should transition to an *objective* standard for whether platforms are moderating in such a way that they should get some relief for the posts they do happen to miss. Platforms big and small collect data on content they moderate[140], so that data could be repurposed to create an objective metric for the success or failure of the platform's moderation. Perfection is impossible, and the effectiveness of approaches may wax and wane, but if, very roughly speaking, 99.5% of harmful content is removed within twenty-four hours of it being posted,[141] then that platform should qualify for protection for liability. Otherwise, they would be subject to a reasonable amount in fines based on a multitude of factors, such as the nature of the content, the harm caused, and the contribution of the content to public discourse. This way, companies that are loosely moderating and catching very little harmful content are not automatically given protection, and all companies are given a concrete goal to strive for in improving their content moderation systems.

138. In 2018, it was estimated that 2.5 quintillion bytes of data were created each day. Bernard Marr, *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*, FORBES (May 21, 2018, 12:42 AM), https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=4f61a7bc60ba.

139. Facebook, Google, and Twitter's CEOs suggested this kind of change to Section 230 in a March 25, 2021 appearance before the House Energy and Commerce Committee. Dylan Byers, *Zuckerberg Calls for Changes to Tech's Section 230 Protections*, NBC NEWS (Mar. 24, 2021, 6:44 AM), https://www.nbcnews.com/news/rcna486. Their suggestion additionally called for a proportionality element to the standard, where smaller companies with fewer resources to devote to moderation are held to a lower standard than larger companies.

140. *See, e.g.*, *Rules Enforcement*, TWITTER (Jan. 11, 2021), https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun; Guy Rosen, *Community Standards Enforcement Report*, FACEBOOK (Feb. 11, 2021), https://about.fb.com/news/2021/02/community-standards-enforcement-report-q4-2020.

141. To be sure, 0.5% of all harmful posts making it through filters would still result in thousands, if not millions of posts evading detection. *See* Marr, *supra* note 138. The threshold suggested is simply an example of how such a system might work; in practice, the cutoff point may have to be much higher, or it may have to vary based on the size of a given platform's userbase.

2. Creating a Government Body to Help Standardize Content Guidelines and Platform Enforcement

To avoid a NetzDG situation[142] in which platforms avoid harsh fines by erring on the side of caution and removing broad amounts of content to bring up their metrics, the federal government should additionally cooperate with platforms to provide independent oversight for the review process and ensure that policies are fair and consistently enforced. This can be done either through an independent federal agency or the judiciary. While this involvement should avoid going as far as enforcement of state-created rules or guidelines, like what the UK suggested in its white paper,[143] this body should at minimum (1) help platforms standardize their rules as to what content is acceptable, and (2) provide an arbitration process for content moderation decisions that users believe are incorrect. Facebook implemented an independent review board in 2020,[144] which operates similarly to how an appellate court might hear a discretionary appeal.[145] Some praise it for setting new precedents for how platforms can approach self-moderation,[146] and others have raised concerns about the scope of its review and the extent of its actual authority over Facebook decisions and policy.[147] The government could look to build further on this concept and design its own body for appeals with input from platforms and industry experts. Alternatively, the government could work in tandem with the Oversight Board and its future counterparts at other platforms by providing them with the necessary guidance and resources to tackle the amount of content they are required to moderate daily.

3. Increasing Individual Disincentives for Disinformation

Because only regulating platforms does not address the core issue of malicious netizens, user-focused laws will still be a necessary evil in a holistic and effective content moderation solution. Currently, federal law provides for fines and prison terms for those who spread false information, but only on very specific topics.[148] These penalties are also increased significantly if the

142. *See supra* Subpart IV.A.

143. WRIGHT & JAVID, *supra* note 90.

144. *Oversight Board Charter*, OVERSIGHT BD., https://www.oversightboard.com/governance (last visited Jan. 24, 2022).

145. Users can appeal Facebook's decisions to the board, submit materials to argue its case to the board, and then a five-member panel will vote and provide a written opinion on the decision they reach. Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418, 2469–74 (2020).

146. *Id.* at 2418 ("Ultimately, the Feature concludes that the Facebook Oversight Board has great potential to set new precedent for user participation in private platforms' governance and a user right to procedure in content moderation.").

147. Taylor Hatmaker, *Facebook's Controversial Oversight Board Starts Reviewing Content Moderation Cases*, TECHCRUNCH (Oct. 22, 2020), https://techcrunch.com/2020/10/22/facebook-oversight-board-controversy.

148. 18 U.S.C. § 35 (1994).

information leads to serious bodily injury or death.[149] However, there are no applicable laws for the more common types of disinformation seen on online platforms today, like political or medical disinformation. Therefore, Congress should look to create an analog to these statutes for those types of disinformation but should take care to limit its application only to situations in which the information leads to injury or death. A statute too broad in its potential application, like the one found in Singapore,[150] could be dangerous in the hands of government officials who could abuse the statute to suppress opposition viewpoints. Limiting it scope to only cases where injury or death occurs would thereby require a minimum threshold result to take place before the statute could be invoked, preventing abuse.

4. Implementing Education Programs through a Cross-Sector Push

It seems from the results abroad that education may be the most effective weapon against disinformation, while simultaneously being least intrusive on First Amendment rights. Media literacy programs for students have been sporadically introduced throughout several cities and states,[151] and there has been a recent push for state legislation requiring media literacy to be incorporated into the curriculum,[152] but much more can still be done. School-age Americans, or those under eighteen, comprise only a quarter of the population,[153] leaving the other three-quarters without much in terms of media literacy resources. Many of the victims of disinformation are typically older and did not grow up with the internet as a core part of their lives.[154] Providing more resources for this large portion of the population to learn media literacy skills could be an important step towards solving the issue.

To address this gap, there should be a cross-sector push, like those in the Nordic countries, to bring these programs to as many people as possible. Non-profits could establish programs to help reach those who are older, or employers could run training sessions for their employees. Traditional media outlets could make use of their reach to send out public services announcements or promote media literacy programs to a broader audience. Full-population media literacy may take decades to achieve, but any small step taken toward increasing the percentage of the media-literate population is a step towards reducing the effects and harms of disinformation campaigns.

---

149. 18 U.S.C. § 1038 (2006).

150. Protection from Online Falsehoods and Manipulation Act, No. 18 of 2019, pt. 3 s 10 (Sing).

151. Alina Tugend, *These Students Are Learning About Fake News and How to Spot It*, N.Y. TIMES (Feb. 20, 2020), https://www.nytimes.com/2020/02/20/education/learning/news-literacy-2016-election.html.

152. MEDIA LITERACY NOW, U.S. MEDIA LITERACY POLICY REPORT 2020 6 (2020), https://medialiteracynow.org/wp-content/uploads/2020/01/U.S.-Media-Literacy-Policy-Report-2020.pdf.

153. *QuickFacts*, U.S. CENSUS BUREAU, https://www.census.gov/quickfacts/fact/table/US/AGE2951#AGE295219 (last visited Jan. 24, 2022).

154. Alexa Lardieri, *Older People More Susceptible to Fake News, More Likely to Share It*, U.S. NEWS (Jan. 9, 2019), https://www.usnews.com/news/politics/articles/2019-01-09/study-older-people-are-more-susceptible-to-fake-news-more-likely-to-share-it.

B. PLATFORM-ENACTED SOLUTIONS

Action by the government does not necessarily mean that platforms can sit back and do nothing. Because the above approaches opt for only a baseline level of control and regulation over how platforms choose to operate, it will also be important for platforms themselves to develop effective systems to moderate their content so they can keep their users safe and qualify for their immunity.

The internet was originally seen as a place where people, no matter how far away, could come together to form communities with each other. But today, companies that run the largest online platforms seem to act more like quasi-governments than they do tools to help users create communities. They are the judge, jury, and executioner of their own platforms: they create their own rules and guidelines that users must follow, make internal content moderation decisions on whether users have violated those rules, and then carry out punishment themselves. For most of the larger platforms, nothing keeps them accountable for their moderation actions beyond the occasional social outrage and cycle of bad press. Users are completely locked out of the decision-making process, and often have no idea what is going on internally at the company. It is no surprise that some groups feel like the biggest tech companies are actively working to stifle certain viewpoints given the lack of clarity surrounding how platform rules are enforced.[155]

1. Designing Platform Content Guidelines Around Their Communities

To remedy these issues of size and transparency, we should look to rebuild platforms and their rules from the ground up. If users can be likened to the citizens, and the platforms to the government, then platforms could be designed to provide users with a say in what the rules and guidelines the platform enforces are. After all, *community* standards or guidelines, by definition, must be enacted based on the desires of the *community* itself. In doing this, community standards should be specific, and vary from location to location and group to group. Mark Zuckerberg envisioned the Oversight Board to "reflect[] the social norms and values of people all around the world," but critics argue that there can be no such thing as a global set of shared values that can govern across all online content.[156]

Reddit approaches this issue by returning to an older system not unlike American federalism. Rather than providing users with one huge, messily organized ocean of information that they can wade through, the site's content is all divided into user-created communities (called "subreddits") that each house discussion on a specific topic, like a sport or a political affiliation. While there are general commonsense sitewide rules that all users must follow (such as respecting other users' privacy or refraining from posting illegal content),[157] each subreddit's members are free to create and enforce their own rules tailored

155. *See* Chappell & Tsioulcas, *supra* note 52.

156. Klonick, *supra* note 145, at 2474–75.

157. *Reddit Content Policy*, REDDIT (Apr. 15, 2021), https://www.redditinc.com/policies/content-policy.

specifically for their content.[158] This brings some transparency[159] to the moderation process, allowing users to feel like they have a part in the system and are in control of their own communities. An analog to this could be implemented at all kinds of online platforms, from new social media to old-school forums. For example, Twitter recently announced the implementation of its own "Birdwatch" program, which would allow users to participate in identifying and countering misinformation by empowering them to identify information in Tweets and add context in the form of supplemental information.[160]

An alternative, or perhaps a supplement, to dividing existing platforms into smaller subcommunities would be to encourage the growth of small or medium-sized platforms. Large platforms like Facebook or YouTube benefit immensely from the network effect, whereby their large user base makes them the "place to be," attracting more users that want to be on the same platforms as people they want to connect with.[161] But this comes with two main costs. First, users join Facebook not because it has tools that no other social media platform does, but because they often feel like they have no other option to avoid being left out of the conversation.[162] This immense population makes large platforms prime targets for abuse by malicious individuals or foreign social media campaigns that want to influence as large of an audience as possible. Second, community moderation on these large platforms can often be difficult to encourage, because the average user lacks any incentive to volunteer their time to help a multibillion-dollar corporation solve its issues.[163]

A natural solution to these problems is to encourage user migration to smaller social media platforms. When online discourse is happening not between faceless names on the internet, but between people in close-knit social groups, there can be more incentives for users to abide by rules of conduct and keep others on the platform accountable. While these small platforms could never fully make a Facebook-sized platform completely obsolete, they are not designed with that goal in mind. Instead, they cater to specific niches and groups by altering their scope and purpose. One example of this is Front Porch Forum, a hyperlocal social media platform that limits membership and discussion to

158. *Reddit 101*, REDDIT (Nov. 22, 2016), https://www.reddit.com/wiki/reddit_101 ("Don't think of Reddit as one giant community. This site is made up of 'sub'reddits, which are all their own communities. Every single post you see on this site belongs to its own community, with its own set of users, and with its own set of rules.").

159. Some have argued that this system still does not bring enough transparency to the platform. Prerna Juneja, Deepika Rama Subramanian & Tanushree Mitra, *Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices*, 4 PROCS. ACM ON HUM.-COMPUT. INTERACTION, no. 4, 2020, at 1 ("Our results reveal a lack of transparency in moderation practices.").

160. Keith Coleman, *Introducing Birdwatch, a Community-Based Approach to Misinformation*, TWITTER: BLOG (Jan. 25, 2021), https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

161. Caroline Banton, *Network Effect*, INVESTOPEDIA (Apr. 3, 2021), https://www.investopedia.com/terms/n/network-effect.asp.

162. Venkatesh, *supra* note 55, at 21:55.

163. *Id.* at 26:23.

residents of the same neighborhoods in Vermont, parts of New York state, and one town in Massachusetts and New Hampshire each.[164] In this more intimate setting, where all the users live only minutes away from each other, everyone has a vested interest in keeping discussion in their neighborhoods civil and healthy. It also allows for more localized moderation that adheres to the norms of the participating neighborhoods.[165]

2. Using Algorithms Alongside Community Moderation

The use of algorithms, especially for moderation, will also be an important issue for platforms going forward. As mentioned earlier, there are myriad issues with the current implementations of automated moderation on social media platforms.[166] However, having some flaws does not mean they should be abandoned completely; they are still a powerful tool that allows for much of the harmful content uploaded online to be stopped before anyone is harmed by it. Instead, the issue with algorithms stems from their current elevation to a status where they are regarded as the solution to all problems that might arise in content moderation.[167] With that as a basis, platforms can use algorithms as an excuse to defend their decisions as scientifically impartial.[168]

Unfortunately, belief in a supreme moderation system that works in mysterious ways and engaging in blind adherence to its decisions abandons critical thinking in favor of technophilia. Algorithms should instead be viewed as tools that serve human ends, rather than ones that rule over our discourse. If platforms truly want to play a part in creating safe communities and giving users voices on the internet, they need to both clarify exactly how their algorithms work and give users at least some input into how the systems are designed. Companies are understandably hesitant to do this, as it would involve being very open with how proprietary technologies core to their business models work. In some cases, complete transparency may be impossible as machine learning creates algorithms that function in ways that are increasingly beyond human

164. *Where is Front Porch Forum Available?*, FRONT PORCH F., https://frontporchforum.com/about-us/service-area (last visited Jan. 24, 2022).

165. *See* Venkatesh, *supra* note 55, at 27:55. Front Porch Forum employs "professional online community managers," but allows neighborhoods themselves to dictate the boundaries of conversation. *See Is FPF for Me?*, FRONT PORCH F., https://frontporchforum.com/isfpfforme (under "Is FPF moderated?") (last visited Jan. 24, 2022).

166. *See supra* Subpart III.A (discussing some of the issues with the current implementation of moderation algorithms).

167. Sarah Jeong, *AI is an Excuse for Facebook to Keep Messing Up*, THE VERGE (Apr. 13, 2018, 2:41 PM), https://www.theverge.com/2018/4/13/17235042/facebook-mark-zuckerberg-ai-artificial-intelligence-excuse-congress-hearings ("Mark Zuckerberg dodged question after question by citing the power of artificial intelligence. Moderating hate speech? AI will fix it. Terrorist content and recruitment? AI again. Fake accounts? AI. Russian misinformation? AI. Racially discriminatory ads? AI. Security? AI.").

168. Gorwa et al., *supra* note 48, at 12 ("[A]utomation is associated with a 'scientific' impartiality that is inherently attractive to platform companies, one that additionally lets them keep their decisions 'non-negotiable' and hidden from view.") (internal citations omitted).

comprehension.[169] However, disclosure does not necessarily have to include any code—the general design framework would be enough of a starting point. For example, what problems was the system designed to solve? What did the designers identify as key contributors to that problem? What does successful implementation look like? By giving users both insight and input into how these questions are answered, algorithms can be adjusted in ways to better suit the needs of a platform's users.

CONCLUSION

While the internet has grown immensely, it has also evolved beyond its original role as a tool to let people communicate and access information. The present laissez-faire system of internet governance has its merits, but it is flawed operating on its own. The time has come for the government to step up and correct some of the major issues plaguing online platforms today. This correction would not and should not result in a state monopoly over the internet. With slight changes to how platforms are held liable for user-generated content, stronger deterrents for the creation of dangerous disinformation, and a nationwide cross-sector push towards media literacy, these issues could be remedied without significantly inhibiting our most foundational principle of speech.

169. Dallas Card, *The "Black Box" Metaphor in Machine Learning*, TOWARDS DATA SCI. (July 5, 2017), https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0.

***