
Fact and Friction: A Case Study in the Fight Against False News

Ayelet Gordon-Tapiero,^{†*} Paul Ohm^{**} & Ashwin Ramaswami^{***}

There is growing recognition within the technology industry of the power of friction to promote important human values and address online harms. Leading technology platforms have begun to slow down communications to give time and space for user thought and deliberation to try to help solve seemingly intractable problems that have emerged from their frictionless designs. Legal scholars have begun to study these uses of friction as a promising self-regulatory and regulatory approach. This Article contributes a new foundation stone to that emerging literature.

Some platforms have injected friction into their social media and messaging services to try to limit the spread of false news. To this end, WhatsApp tags some messages as “frequently forwarded” and places limits on how those messages can be shared. The stakes are high, as the past several years have seen a rise in the prevalence of false news spread via WhatsApp, contributing to election disruptions, mob violence, and even deaths.

[†] Copyright © 2023 Ayelet Gordon-Tapiero, Paul Ohm & Ashwin Ramaswami.

^{*} Postdoctoral Fritz Research Fellow, Initiative on Tech & Society, Georgetown University.

^{**} Professor of Law, Georgetown University Law Center.

^{***} JD Candidate, 2024, Georgetown University Law Center. The authors would like to thank Ryan Calo, David Clark, Elizabeth Edenberg, Brett Frischmann, Woody Hartzog, Marcia Hofmann, Meg Jones, William Lehr, Katrina Ligett, Kobbi Nissim, Eric Null, Blake Reid, Nathan Reiting, Alexis Shore, Ashkan Soltani, Harry Surden, Wayne Unger, Rebecca, Wexler, Christopher Yoo, and the participants of the Privacy Law Scholars Conference; the Law and CS Roundtable at the Center for Technology, Innovation & Competition, at the University of Pennsylvania; and the Technology Law Scholars colloquium at Georgetown Law. This work has been supported by the Fritz Family Fellowship Program and the Georgetown McCourt School of Public Policy’s Tech & Public Policy Program.

WhatsApp's approach has been widely hailed and held out as a best practice by policymakers, journalists, and academics alike, yet it has never been the subject of an extensive technical or policy analysis until now. This Article reports the results of analysis conducted by a team of legal and technical scholars. While we confirmed many of WhatsApp's claims about its mechanisms, we were also able to find many ways to circumvent their celebrated restrictions.

Today's self-regulatory attempts to introduce friction-in-design can serve as a blueprint for tomorrow's regulatory mandates. From the insights generated by our technical analysis we draw lessons for policymakers considering new laws mandating WhatsApp-style friction into social media and messaging services and other forms of friction-in-design regulation.

TABLE OF CONTENTS

INTRODUCTION.....	173
I. FRICTION	180
A. <i>Designing for a Frictionless Experience</i>	181
B. <i>The Rise of Friction-in-Design</i>	183
C. <i>Types of Friction</i>	185
1. Friction Offline	186
2. Friction Online.....	187
II. FALSE NEWS.....	191
A. <i>The Harms of False News</i>	192
1. The Societal Nature of the Harms.....	193
2. Terminology	196
B. <i>The Far and Wide Reach of False News</i>	198
1. Who Believes False News?.....	198
2. Who Spreads False News?	201
III. FRICTION AS A TOOL TO LIMIT THE SPREAD OF FALSE NEWS.....	202
A. <i>WhatsApp — Overview</i>	203
B. <i>Limiting Forwarding as a Way to Fight False News</i>	205
1. False News on WhatsApp	205
2. WhatsApp Limits Forwards to Combat False News....	207
3. The World Responds Positively.....	211
4. The Surprising Amount We Do Not Know.....	214
IV. TECHNICAL INVESTIGATION OF WHATSAPP'S FRICTION.....	216
A. <i>Methodology</i>	216

B.	<i>Behavioral Analysis</i>	218
1.	Behavioral Analysis Explained	219
2.	Behavioral Analysis Results: The User’s Point of View	220
C.	<i>Static Analysis of Source Code</i>	225
1.	Static Analysis	225
2.	Static Analysis Results.....	227
D.	<i>Dynamic Analysis of Running Source Code</i>	228
1.	Dynamic Analysis.....	228
2.	Dynamic Analysis Results	229
E.	<i>Why Circumvention Matters</i>	232
1.	Does Circumvention Matter?.....	232
2.	Is Anyone Circumventing These Restrictions?	232
V.	GUIDELINES FOR POLICYMAKERS MANDATING FRICTION	235
A.	<i>Code Is Law; Law Should Be (Based on an Understanding of) Code</i>	238
B.	<i>Content-Neutrality</i>	239
C.	<i>Does It Really Work?</i>	242
1.	On the Importance of Being Measured	242
2.	What Type of User Is Impacted.....	243
D.	<i>It’s a Never-Ending Story</i>	246
1.	Tunability.....	246
2.	Friction Will Ignite an Arms Race	247
E.	<i>Friction Can Be Used as a Temporary Fix</i>	249
	CONCLUSION	250

INTRODUCTION

Regulators around the world have been urged to implement new forms of regulation to combat false news.¹ These calls are of increasing

¹ See Rebecca K. Helm & Hitoshi Nasu, *Regulatory Responses to “Fake News” and Freedom of Expression: Normative and Empirical Evaluation*, 21 HUM. RTS. L. REV. 302, 302-03 (2021); Irini Katsirea, “Fake News”: *Reconsidering the Value of Untruthful Expression in the Face of Regulatory Uncertainty*, 10 J. MEDIA L. 159, 159-60 (2018); Chris Marsden, Trisha Meyer & Ian Brown, *Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?*, 36 COMPUT. L. & SEC. REV. 1, 3-6 (2020); Andrei Richter, *Fake News and Freedom of the Media*, 8 J. INT’L MEDIA & ENT. L. 1, 10-14 (2019); Fredrick Wilson & Muhammad A. Umar, *The Effect of Fake News on Nigeria’s Democracy Within the*

importance as false news sows division, impedes dialogue, inspires violence, and interferes with free democratic elections.² Current approaches and proposals to address the crisis of false news do little more than tinker with the insufficient privately run content moderation systems already in place, while calls for government to do more are attacked as impermissible infringements on free expression.³ A better, new path forward is for regulators to draw inspiration from recent advances in building *friction* into communications platforms.

Despite the technology industry's well-documented veneration for frictionless design,⁴ recently some of the world's largest platforms have begun purposely inserting friction into their products and services not only to fight false news but to advance many goals.⁵ Often the friction is aimed at promoting an important human value beyond simple efficiency or profit maximization, such as to advance fairness, trust, or consensus.⁶

Premise of Freedom of Expression, 17 GLOB. MEDIA J. 1, 10 (2019); Daniela C. Manzi, Note, *Managing the Misinformation Marketplace: The First Amendment and the Fight Against Fake News*, 87 FORDHAM L. REV. 2623, 2637-48 (2019). For a discussion on our use of the phrase, "false news," see *infra* notes 81 and 109.

² Simone Chambers, *Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere?*, 69 POL. STUD. 147, 149-51 (2021); Rachael Craufurd Smith, *Fake News, French Law and Democratic Legitimacy: Lessons for the United Kingdom?*, 11 J. MEDIA L. 52, 63-64 (2019); Deen Freelon & Chris Wells, *Disinformation as Political Communication*, 37 POL. COMMUN. 145, 146-47 (2020); Linda Monsees, *Information Disorder, Fake News and the Future of Democracy*, 20 GLOBALIZATIONS 153, 156-58 (2023); Susan Morgan, Interview, *Fake News, Disinformation, Manipulation and Online Tactics to Undermine Democracy*, 3 J. CYBER POL'Y 39, 39-40 (2018).

³ Evelyn Douek, *Content Moderation as Systems Thinking*, 136 Harv. L. Rev. 526, 535-39, 557 (2022); Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 Yale L.J. 2418, 2427-39 (2020); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598, 1603-14 (2018) [hereinafter Klonick, *The New Governors*].

⁴ Ellen P. Goodman, *Digital Fidelity and Friction*, 21 NEV. L.J. 623, 646-48 (2021); William McGeeveran, *The Law of Friction*, 2013 U. CHI. LEGAL F. 15, 19-21 (2013).

⁵ Paul Ohm & Jonathan Frankle, *Desirable Inefficiency*, 70 FLA. L. REV. 777, 790-97 (2018).

⁶ *Id.* at 797-98.

Since code is law,⁷ these emerging examples of friction voluntarily integrated into technology should be viewed as a form of underappreciated self-regulation.⁸ Today's self-regulation can open new pathways for tomorrow's regulation. This Article explores whether and how regulators might mandate or incentivize new uses of friction to tackle false news. It contributes a foundation stone for the emerging literature around friction-in-design.⁹

The frictionless design of social media and other communications platforms has exacerbated the spread of false news.¹⁰ Research has found that people share false news further and faster than true news.¹¹ Part of the reason is that people do not stop to assess the reliability and value of content before clicking "share."¹² Recognizing that their design decisions have contributed to the problem, some platforms have started

⁷ LAWRENCE LESSIG, CODE: VERSION 2.0, at 1-8 (2006); see Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 568-69 (1998).

⁸ David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts & Jonathan L. Zittrain, *The Science of Fake News*, 359 SCIENCE 1094, 1095 (2018).

⁹ See, e.g., Brett Frischmann & Susan Benesch, *Friction-In-Design Regulation as 21st Century Time, Place, and Manner Restriction*, 25 YALE J.L. & TECH. 376, 377 (2023) (discussing how a "friction-in-design" approach could be seen as a form of time, place, and manner regulation under First Amendment jurisprudence); Goodman, *supra* note 4, at 624-45 (framing disclosure requirements as a form of friction); McGeeveran, *supra* note 4, at 17 (highlighting the benefits of friction in data sharing); Ohm & Frankle, *supra* note 5, at 785 (demonstrating that a certain level of friction can lead to desirable results); see also Brett Frischmann & Paul Ohm, *Governance Seams*, 37 HARV. J.L. & TECH. (manuscript at 12) (forthcoming 2023).

¹⁰ Lazer et al., *supra* note 8, at 1096.

¹¹ Soroush Vosoughi, Deb Roy & Sinan Aral, *The Spread of True and False News Online*, 359 SCIENCE 1146, 1146 (2018). Craig Silverman from BuzzFeed News reported a similar trend, finding that "the most popular . . . fabricated stories were shared more widely than the most popular stories from mainstream media." CLAIRE WARDLE & HOSSEIN DERAKHSHAN, INFORMATION DISORDER: TOWARD AN INTERDISCIPLINARY FRAMEWORK FOR RESEARCH AND POLICY MAKING 11 (2017), <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c> [<https://perma.cc/WZW2-4ZF8>].

¹² Gordon Pennycook & David G. Rand, *The Psychology of Fake News*, 25 TRENDS COGNITIVE SCI. 388, 393 (2021).

slowing things down. Perhaps the most widely noted and celebrated example is WhatsApp's introduction, over the past five years, of limits on its users' ability to forward messages quickly and at scale as an express attempt to limit the virality of false news.¹³

WhatsApp has been used as a tool to spread false news all around the world.¹⁴ In some countries, elections have even been dubbed "WhatsApp Elections" due to the significant role the spread of false news on WhatsApp played in them.¹⁵ Journalists and researchers have documented the role WhatsApp played in spreading false news in the period leading up to the 2018 presidential elections in Brazil,¹⁶ and in

¹³ Philippe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro O.S. Vaz de Melo & Fabrício Benevenuto, *Can WhatsApp Counter Misinformation by Limiting Message Forwarding?*, in 1 COMPLEX NETWORKS AND THEIR APPLICATIONS VIII, at 372, 372, 378 (Hocine Cherifi, Sabrina Gaito, José Fernando Mendes, Esteban Moro & Luis Mateus Rocha eds., 2019).

¹⁴ Melo et al., *supra* note 13, at 372; see Rita El Khoury, *WhatsApp Now Lets You Share and Forward a Message to Multiple Chats (with Frequent Chats on Top)*, ANDROID POLICE (Aug. 11, 2016), <https://www.androidpolice.com/2016/08/11/whatsapp-now-lets-forward-share-message-multiple-chats-easily-displays/> [<https://perma.cc/DKF6-E9TQ>].

¹⁵ See, e.g., Nic Cheeseman, Jonathan Fisher, Idayat Hassan & Jamie Hitchen, *Social Media Disruption: Nigeria's WhatsApp Politics*, 31 J. DEMOCRACY 145, 153 (2020) (analyzing WhatsApp's role in the Nigerian elections); Priyanjana Bengani, *India Had Its First "WhatsApp Election." We Have a Million Messages from It*, COLUM. JOURNALISM REV. (Oct. 16, 2019), https://www.cjr.org/tow_center/india-whatsapp-analysis-election-security.php [<https://perma.cc/9V8F-WLZ5>] (describing WhatsApp's role in the Indian elections); Jamie Hitchen, Jonathan Fisher, Nic Cheeseman & Idayat Hassan, *How WhatsApp Influenced Nigeria's Recent Election — And What It Taught Us About "Fake News,"* WASH. POST (Feb. 15, 2019, 6:00 AM EST), <https://www.washingtonpost.com/news/monkey-cage/wp/2019/02/15/its-nigerias-first-whatsapp-election-heres-what-were-learning-about-how-fake-news-spreads/> [<https://perma.cc/B64R-69KY>] (highlighting that the impact of WhatsApp on Nigerian elections was similar to its influence in Brazil); Madhumita Murgia, Stephanie Findlay & Andres Schipani, *India: The WhatsApp Election*, FIN. TIMES (May 4, 2019), <https://www.ft.com/content/9fe88fba-6cod-11e9-a9a5-351eeaf6d84> [<https://perma.cc/FL8L-HZ62>] (describing WhatsApp's role in Indian elections).

¹⁶ Daniel Avelar, *WhatsApp Fake News During Brazil Election "Favoured Bolsonaro,"* GUARDIAN (Oct. 30, 2019, 12:00 PM EDT), <https://www.theguardian.com/world/2019/oct/30/whatsapp-fake-news-brazil-election-favoured-jair-bolsonaro-analysis-suggests> [<https://perma.cc/E3BU-GVRK>]; Matheus Magenta, Juliana Gragnani & Felipe Souza, *How WhatsApp Is Being Abused in Brazil's Elections*, BBC NEWS (Oct. 24, 2018), <https://www.bbc.com/news/technology-45956557> [<https://perma.cc/2C7Z-Y6U5>].

elections in India and Nigeria.¹⁷ The problem extends beyond elections, as WhatsApp has also been used to spread conspiracy theories and to propagate false information about COVID-19 and the alleged effects of vaccines.¹⁸ To combat false news, starting in 2018, WhatsApp has implemented successive waves of restrictions on the number of chats to which a user could forward content¹⁹ and has introduced user interface features to alert users that they were receiving content forwarded by others.²⁰ Journalists, policymakers, activists, and scholars have hailed these changes as bold and effective means for limiting the spread of false news.²¹

Although the WhatsApp example has been widely noted, its technological implementation has been under analyzed and, as we have learned, somewhat misunderstood. This article undertakes a rigorous technical dissection of the measures WhatsApp has put in place to combat forwarded message propagation, subjecting the single most prominent example of friction-in-design to an overdue examination.²²

¹⁷ Kiran Garimella & Dean Eckles, *Images and Misinformation in Political Groups: Evidence from WhatsApp in India*, 1 Harv. Kennedy Sch. Misinformation Rev. 1 (2020); Kevin Ponniah, *WhatsApp: The “Black Hole” of Fake News in India’s Election*, BBC News (Apr. 6, 2019), <https://www.bbc.com/news/world-asia-india-47797151> [<https://perma.cc/7H4B-VQCN>]; Philippa Williams, *Technology Could Torpedo India’s First WhatsApp Election*, Yahoo! Fin. (Mar. 4, 2019, 2:27 AM), <https://finance.yahoo.com/news/technology-could-torpedo-india-first-072702633.html> [<https://perma.cc/2NQB-4LK9>].

¹⁸ Helm & Nasu, *supra* note 1, at 304; Mozer de Miranda Ramos, Rodrigo de Oliveira Machado & Elder Cerqueira-Santos, “It’s True! I Saw It on WhatsApp:” *Social Media, Covid-19, and Political-Ideological Orientation in Brazil*, 30 TRENDS PSYCH. 570, 581 (2022); Tiffany Hsu, *As Covid-19 Continues to Spread, So Does Misinformation About It*, N.Y. TIMES, <https://www.nytimes.com/2022/12/28/technology/covid-misinformation-online.html> (last updated Jan. 1, 2023) [<https://perma.cc/2RPE-2P94>]; Tony Romm, *Fake Cures and Other Coronavirus Conspiracy Theories Are Flooding WhatsApp, Leaving Governments and Users with a “Sense of Panic,”* WASH. POST (Mar. 2, 2020, 10:58 AM EST), <https://www.washingtonpost.com/technology/2020/03/02/whatsapp-coronavirus-misinformation/> [<https://perma.cc/6UU2-5Q95>].

¹⁹ *About Forwarding Limits*, WHATSAPP HELP CTR., https://faq.whatsapp.com/1053543185312573/?locale=en_US (last visited July 4, 2023) [<https://perma.cc/3SGT-3UR9>] [hereinafter *About Forwarding Limits*].

²⁰ *Id.*

²¹ See *infra* Part III.B.3 (summarizing praise for WhatsApp forwarding restrictions).

²² See Frischmann & Benesch, *supra* note 9, at 413.

Our results belie the widely held idea that WhatsApp's measures are foolproof and thus cast some doubt on the effectiveness of the approach.

Specifically, we reveal that WhatsApp's forwarding restrictions are easy to identify and circumvent. By carefully examining the underlying code of the app — using techniques common in computer security research but less common in prior legal scholarship — we were able to circumvent both WhatsApp's mechanism for tracking how many times a message had been forwarded and the restrictions imposed on forwarded messages, with these interventions requiring only a modest level of technical know-how. Once we knew what to look for, we found several real examples of third-party apps that ignored or circumvented these controls by searching through online source code repositories. It would be very easy to develop an app or browser extension that any non-technical user could install to evade all of WhatsApp's forwarding restrictions. Consistent with responsible disclosure guidelines, we have shared our findings with WhatsApp prior to publication.²³

Our observations can guide regulators considering new mandates to integrate friction into technology. First, the WhatsApp example highlights an important relationship between friction-based regulation and First Amendment law, building on recent work by Brett Frischmann and Susan Benesch.²⁴ Unlike many other proposals to use regulation to combat false news, WhatsApp's approach is content-neutral. All messages are subject to the forwarding limitations, regardless of type, content, truth or falsity, or viewpoint.²⁵ We believe content-neutrality is

²³ For best practices around vulnerability disclosure, see, for example, NAT'L TELECOMMS. INFO. ADMIN. AWARENESS & ADOPTION GRP., VULNERABILITY DISCLOSURE ATTITUDES AND ACTIONS (2016), https://www.ntia.doc.gov/files/ntia/publications/2016_ntia_a_a_vulnerability_disclosure_insights_report.pdf [<https://perma.cc/8B24-V57K>]; *Coordinated Vulnerability Disclosure Process*, CYBERSEC. & INFRASTRUCTURE SEC. AGENCY, <https://www.cisa.gov/coordinated-vulnerability-disclosure-process> (last visited July 4, 2023) [<https://perma.cc/57U8-RS5V>]; Nancy Gariché, *Coordinated Vulnerability Disclosure (CVD) for Open Source Projects*, GITHUB BLOG (Feb. 9, 2022), <https://github.blog/2022-02-09-coordinated-vulnerability-disclosure-cvd-open-source-projects/> [<https://perma.cc/2NND-X3SN>]; *Vulnerability Disclosure Cheat Sheet*, OPEN WORLDWIDE APPLICATION SEC. PROJECT CHEAT SHEET SERIES, https://cheatsheetseries.owasp.org/cheatsheets/Vulnerability_Disclosure_Cheat_Sheet.html (last visited July 4, 2023) [<https://perma.cc/8QYU-XUZ9>].

²⁴ Frischmann & Benesch, *supra* note 9, at 413.

²⁵ *About Forwarding Limits*, *supra* note 19.

a common feature of frictional approaches, giving us hope that friction-in-design solutions may often withstand First Amendment scrutiny.

Second, our analysis highlights the different effects friction has on different types of users. While the type of friction implemented by WhatsApp may restrict ordinary users unknowingly engaged in spreading misinformation, it is not likely to have much impact on highly motivated “superusers” who purposely create and spread disinformation.²⁶ This is another characteristic feature of friction. The decision to integrate friction is often the opening shot in an arms race, as motivated and skilled users attempt to overcome any level of friction introduced.²⁷ Platforms and policymakers need to decide how far they want to engage in this arms race.

Third, unlike other regulatory approaches, most friction approaches are *tunable*. Platforms and policymakers can decide how much friction to apply, and they can dial that amount up or down in response to changing circumstances and the different types of users they are looking to slow down.

On an analytical level, this Article argues that effective policy regulating the operation of technology must be informed by a robust technical understanding.²⁸ On a normative level, we recommend using friction as a tool to counter the spread of false news. From a constitutional perspective, we explain why content-neutral friction is likely to withstand First Amendment scrutiny. Finally, on a practical

²⁶ Paul Ohm, *The Myth of the Superuser: Fear, Risk, and Harm Online*, 41 UC DAVIS L. REV. 1327, 1396-99 (2008).

²⁷ Lee Kovarsky, *A Technological Theory of the Arms Race*, 81 IND. L.J. 917, 937-39 (2006).

²⁸ See Aloni Cohen & Sunoo Park, *Compelled Decryption and the Fifth Amendment: Exploring the Technical Boundaries*, 32 HARV. J.L. & TECH. 169, 173 (2018); Aloni Cohen & Kobbi Nissim, *Towards Formalizing the GDPR’s Notion of Singling Out*, 117 PROC. NAT’L. ACAD. SCI. 8344, 8344 (2020); Ayelet Gordon-Tapiero, Alexandra Wood & Katrina Ligett, *The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization*, 25 VAND. J. ENT. & TECH. L. 635, 636 (2023); Micah Altman, Aloni Cohen, Francesca Falzon, Evangelia Anna Markatou, Kobbi Nissim, Michel José Reymond, Sidhant Saraogi & Alexandra Wood, *A Principled Approach to Defining Anonymization as Applied to EU Data Protection Law 3-5* (May 10, 2022) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4104748 [<https://perma.cc/WK46-P8YC>].

level, we highlight challenges likely to be encountered by any party integrating friction into technology.

The Article proceeds as follows. Part I describes the technology industry's preference for frictionless technology, as well as its recent rediscovery of how friction can sometimes be used to promote various human values, giving both offline and online examples and surveying the emerging literature studying friction-in-design. Part II presents the dangerous phenomenon of false news, which is harmful to individuals, but even more dangerous to society. In particular, it describes the destructive effect false news has had on the public's trust in democracy as a system of governance, as well as on its institutions and basic processes, such as elections. It identifies the psychological causes leading people to believe and spread false news farther and faster than true news. Part III recounts how WhatsApp has implemented friction in the past few years as a way to combat false news, and it highlights how surprisingly little this move has received academic scrutiny, leaving many open questions about how this friction operates. Part IV presents the original technical analyses we engaged in to better understand WhatsApp's use and implementation of friction. While we were able to confirm WhatsApp's claims about how its friction operated, we were surprised to learn that most of these restrictions were relatively easy to circumvent, and we even unearthed evidence that significant numbers of users are using tools that circumvent some of these constraints. Based on our analysis, Part V identifies policy guidelines for regulators wishing to mandate friction to achieve a human goal.

Friction can be a helpful tool in overcoming harms that may have previously seemed unsurmountable. It is one of the few options available to policymakers seeking to overcome the harms generated by the widespread dissemination of false news. We would be wise to harness the potential that this tool has in protecting our basic values and safeguarding our democratic institutions.

I. FRICTION

Each professional discipline instills its basic dogmas within its new entrants in their very early stages of training. For law students these

include rules such as *res ipsa loquitor*²⁹ and *caveat emptor*.³⁰ Computer science students are also taught their own set of such basic rules. From the early stages of their undergraduate degree, computer science students are taught to optimize for technical efficiency.³¹ They learn to search for, and surmount, any obstacles that might slow down a computational process or make it more cumbersome than absolutely necessary.³² As Ellen Goodman notes, “for the engineer, friction is ‘any sort of irritating obstacle’ to overcome.”³³ The result has been that technology companies aspire to create technology that is as seamless as possible, eliminating any type of friction that would complicate use of the technology along the way.³⁴

A. Designing for a Frictionless Experience

A frictionless experience has become such a widespread goal in the technology industry that consumers have been conditioned to look for it.³⁵ Leading technology companies like Apple, Amazon and Uber have

²⁹ See analysis of the doctrine of *res ipsa loquitor* in Charles E. Carpenter, *The Doctrine of Res Ipsa Loquitor*, 1 U. CHI. L. REV. 519, 519-23 (1934).

³⁰ See analysis of the doctrine of *caveat emptor* in Morton J. Horwitz, *The Historical Foundations of Contract Law*, 87 HARV. L. REV. 917, 945 (1974).

³¹ Ohm & Frankle, *supra* note 5, at 786, 800 (“Software is and will remain the product of human engineering and organization.”).

³² See, e.g., Jim Gao, Machine Learning Applications for Data Center Optimization (2014) (unpublished manuscript), <https://research.google/pubs/pub42542.pdf> [<https://perma.cc/MVG8-B5CV>] (describing the role of optimization algorithms in optimizing data center performance); Larry Hardesty, *Optimizing Optimization Algorithms*, MIT NEWS (Jan. 21, 2015), <https://news.mit.edu/2015/optimizing-optimization-algorithms-0121> [<https://perma.cc/B4DC-L6UK>] (describing the role of optimization algorithms in solving engineering problems).

³³ Goodman, *supra* note 4, at 648 (quoting in part from McGeveran, *supra* note 4, at 51).

³⁴ Frischmann & Benesch, *supra* note 9, at 445-46.

³⁵ SHOSHANA ZUBOFF, *THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER* 242 (2019) (“We have seen the urgency with which surveillance capitalists pursue the elimination of ‘friction’ as a critical success factor in supply operations.”); Future Friendly, *The History of Experience: Accident or Design?*, MEDIUM (June 14, 2016), https://medium.com/@Future_Friendly/the-history-of-experience-accident-or-design-dc35f5a0c465 [<https://perma.cc/GLJ2-WPLQ>]; Aaron Levie, *The Simplicity Thesis*, FAST CO. (May 2, 2012), <https://www.fastcompany.com/1835983/simplicity-thesis> [<https://perma.cc/J57V-FTNJ>]; Christian Terwiesch & Nicolaj Siggelkow, *Designing a Seamless Digital Experience for Customers*, HARV. BUS. REV. (Dec. 2,

been able to translate the minimal friction designed into their products and services into billions of dollars.³⁶ When Facebook CEO Mark Zuckerberg introduced the platform's timeline in 2011, he referred to its features as realizing the promise of "real-time serendipity" in "frictionless experiences."³⁷ "From here on out," he explained, "it's a frictionless experience."³⁸ The decision to prioritize a frictionless user experience has had far-reaching implications on the way technology is designed and on the way users interact with it.³⁹

Several decades ago, Joel Reidenberg and Lawrence Lessig taught us that code is law.⁴⁰ The way technological systems are designed sets the rules governing the activity of their users and those influenced by them.⁴¹ The strong preference for frictionless design was never explicitly imposed upon the tech industry by an outside regulator. Still, the incentive to design frictionless technology is so prevalent that it can be thought of as a self-regulatory standard for technology companies.⁴²

2021), <https://hbr.org/2021/12/designing-a-seamless-digital-experience-for-customers> [<https://perma.cc/Y6MZ-9QYH>].

³⁶ Shubham Agarwal, *Technology Is Easier than Ever to Use — And It's Making Us Miserable*, Digit. Trends (Oct. 25, 2020), <https://www.digitaltrends.com/web/the-frictionless-internet/> [<https://perma.cc/5ME9-4ECP>]; Kevin Roose, *Is Tech Too Easy to Use?*, N.Y. Times (Dec. 12, 2018), <https://www.nytimes.com/2018/12/12/technology/tech-friction-frictionless.html> [<https://perma.cc/NN5F-KG37>].

³⁷ Romchik Nikolaenkov, *F8 2011 Keynote*, YOUTUBE, at 38:30 (Sept. 24, 2011), <https://www.youtube.com/watch?v=9r46UeXCzoU&t=2305s> [<https://perma.cc/952B-5ZYJ>]; Roose, *supra* note 36; Matt Rosoff, *Mark Zuckerberg Rolls Out Big Changes for Facebook*, BUS. INSIDER (Sept. 22, 2011, 10:16 AM PDT), <https://www.businessinsider.com/live-facebooks-big-day-2011-9> [<https://perma.cc/BRA2-RXN6>].

³⁸ Robert Hof, *LIVE with Mark Zuckerberg at F8: Facebook is Your Life*, FORBES (Sep 22, 2011, 01:11 PM EDT), <https://www.forbes.com/sites/roberthof/2011/09/22/live-with-mark-zuckerberg-at-facebook-f8/?sh=3fd8a7e6432e> [<https://perma.cc/DYC3-ZSUU>]; Rosoff, *supra* note 37. In the early 2010s frictionless sharing was heralded as "the wave of the future." McGeeveran, *supra* note 4, at 15. The following year, Facebook rolled back certain aspects of its frictionless sharing, explained at least partially due to negative reactions the feature received. *Id.* at 21-22.

³⁹ Ohm & Frankle, *supra* note 5, at 800 (describing how friction is integrated to promote human values).

⁴⁰ LESSIG, *supra* note 7, at 1; *see* Reidenberg, *supra* note 7, at 568-69.

⁴¹ LESSIG, *supra* note 7, at 7.

⁴² *See* Klonick, *The New Governors*, *supra* note 3, at 1601 (highlighting the self-regulatory nature of platforms' content moderation practices). On self-regulation, *see*

The “regulators” selecting frictionless design are the software engineers designing it, as well as the CEOs, CTOs, CPOs (chief privacy officers), CROs (chief risk officers), legal advisors, product managers, team leaders, security officers, and other actors employed by the leading technology companies that influence the final outcome of a product.⁴³ It is their decisions that govern the activity of the users who operate their systems.⁴⁴

B. *The Rise of Friction-in-Design*

Despite leading technology companies’ basic aspiration for frictionless design, recently a new trend has emerged.⁴⁵ As documented by legal scholars,⁴⁶ technology companies have started to purposely insert friction into their products and services.⁴⁷ One way companies do

Balázs Muraközy & Pál Valentiny, *Alternatives to State Regulation: Self- and Co-Regulation*, in COMPETITION AND REGULATION 54 (Pál Valentiny, Ferenc László Kiss, Krisztina Antal-Pomázi & Csongor István Nagy eds., 2015); Colin Provost, *Governance and Voluntary Regulation*, in THE OXFORD HANDBOOK OF GOVERNANCE 554 (David Levi-Faur ed., 2012).

⁴³ For a fascinating account of the dynamics between these actors and how they influence the privacy practices of technology companies, see generally ARI EZRA WALDMAN, *INDUSTRY UNBOUND: THE INSIDE STORY OF PRIVACY, DATA, AND CORPORATE POWER* (2021).

⁴⁴ See Klonick, *The New Governors*, supra note 3, at 1603. An important legal application utilizing frictionless design involves online contracts, such as privacy policies. The average individual clicks “I agree” on numerous online contracts, without ever having read them. This consent process reflects architecture purposely designed to include as little friction as possible to “minimize transaction costs, maximize efficiency, minimize deliberation, engineer complacency, and, as a result, nudge people to behave like simple machines.” BRETT FRISCHMANN & EVAN SELINGER, *RE-ENGINEERING HUMANITY* 68-69, 288 (2018).

⁴⁵ Goodman, supra note 4, at 649 (“[P]latforms themselves are voluntarily moving to implement frictive solutions.”); Ohm & Frankle, supra note 5, at 785.

⁴⁶ See, e.g., Goodman, supra note 4 (framing disclosure rules in the media as a form of friction); McGeeveran, supra note 4 (detailing the ways in which leading media companies, such as Netflix, integrate friction into their sharing practices); Ohm & Frankle, supra note 5 (analyzing the integration of friction as a form to promote human values); Frischmann & Benesch, supra note 9 (considering the integration of friction by companies like WhatsApp, Apple, and Twitter).

⁴⁷ Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides & Ian Renfree, *Design Frictions for Mindful Interactions: The Case for Microboundaries*, in

this is by encouraging users to slow down. Users are asked: “not so fast, show me you’re human and not a bot.”⁴⁸ Including friction in online activities enables users to “open pathways for reflection,”⁴⁹ encouraging them to purposefully take notice of what they are doing and not just respond automatically.⁵⁰ Users are more likely to do this when they are prompted to take the necessary time to deliberate about their choice before making it, and when they have the tools, ability and “mental space and inclination to raise cognitive defenses.”⁵¹

This type of friction relates to what we know about human behavioral psychology. Daniel Kahneman and Amos Tversky labeled two systems driving human behavior.⁵² System one is the more intuitive, instinctive mechanism. It includes primal instincts such as flight or fight, drawing our hand away from something hot, etc. System two includes more deliberative activities. Moral judgment and deliberative processes occur under system two. Some actions, known as dual tasks,⁵³ can be carried out under both systems, with varying outcomes. Take driving as an example.⁵⁴ An experienced driver often feels as if they are driving automatically, an action governed by their system one thinking. Even an experienced driver, however, when driving on a snowy road in the dark, will carry out their actions more deliberately. They may lean forward, grip the wheel, and turn the music off. An external constraint, in this case the objective conditions of the road, caused the driver to activate

PROCEEDINGS OF THE 2016 CHI CONFERENCE EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS 1389, 1392-94 (2016); Goodman, *supra* note 4, at 650-51.

⁴⁸ See Frischmann & Benesch, *supra* note 9, at 14 n.59.

⁴⁹ Goodman, *supra* note 4, at 624.

⁵⁰ Brett Frischmann, *Here’s Why Tech Companies Abuse Our Data: Because We Let Them*, GUARDIAN (Apr. 10, 2018, 1:00 EDT), <https://www.theguardian.com/commentisfree/2018/apr/10/tech-companies-data-online-transactions-friction> [<https://perma.cc/R9ZP-5HQJ>] (“Friction is resistance; it slows things down. And in our hyper-rich, fast-paced, attention-deprived world, we need opportunities to stop and think, to deliberate and even second-guess ourselves and others. This is how we develop the capacity for self-reflection; how we experiment, learn and develop our own beliefs, tastes and preferences; how we exercise self-determination. This is our free will in action.”).

⁵¹ Goodman, *supra* note 4, at 649.

⁵² DANIEL KAHNEMAN, THINKING, FAST AND SLOW 19 (2011).

⁵³ Daniel Kahneman, *Maps of Bounded Rationality: Psychology for Behavioral Economics*, 93 AM. ECON. REV. 1449, 1451 (2003).

⁵⁴ Ohm & Frankle, *supra* note 5, at 790.

her system two. It is thus possible to manipulate the conditions of the road — think speed bumps — to cause an individual to switch a system one-powered activity to one governed by system two. Introducing friction into an otherwise automatic activity encourages the user to activate their system two thinking, conduct a more deliberative thought process, and act in a more calculated way.⁵⁵ When technologists introduce friction into their systems, they rely on this very insight about the psychology of the human mind.

Although most of the prominent examples of friction to date have been imposed in self-regulatory fashion by the technology industry, we suggest that external regulators consider mandating the integration of friction in some cases by asking questions such as: when is friction particularly helpful? What types of friction are likely to be most effective? And what kinds of psychological challenges can friction overcome?⁵⁶

C. Types of Friction

In their recent article, *Friction-In-Design Regulation as 21st Century Time, Place and Manner Restriction*, legal scholars Brett Frischmann and Susan Benesch identify six parameters that are helpful in evaluating and comparing various instances of friction. These are: (1) the type of friction; (2) the direct effect of friction on subjects; (3) the architectural design of friction; (4) the purpose of friction: the intended impact of

⁵⁵ Laura Hedeem, *When Design Friction Is a Good Thing*, UX COLLECTIVE (Oct. 8, 2020), <https://uxdesign.cc/when-design-friction-is-a-good-thing-d4ab56c4049e> [<https://perma.cc/R7UQ-KRB4>].

⁵⁶ Such a trend can occur either because regulators learn of the value of friction and grow to support it, or because of regulatory capture. For the purpose of this Article, there is not a strong difference between the two incentive structures driving a regulatory requirement to introduce friction as a tool to promote values. George J. Stigler, *The Theory of Economic Regulation*, 2 BELL J. ECON. & MGMT. SCI. 3, 3 (1971). Similarly, the question of the regulatory mechanism used to introduce friction is outside the scope of this article. Three of these tolls may include induced self-regulation. See Julia Black, *Decentering Regulation: Understanding the Role of Regulation and Self-Regulation in a “Post-Regulatory” World*, 54 CURRENT LEGAL PROBS. 103, 129-31 (2001), for an analysis of different types of self-regulation.

friction; (5) the scope of friction; and (6) the governance of the friction.⁵⁷

The six categories are insightful when conducting fine-grained analyses of friction, as well as when comparing different types of friction. In the following part we look at them together and inquire about the intended goals of various types of friction (Frischmann-Benesch parameter four), their implementation (one, three, five) and consequences (two). Additionally, we follow Frischmann and Benesch in adopting a broad perspective asking: who it is that should determine when friction should be used and to promote what values (six)? What actor should be charged with deploying friction and assessing the advantages and costs it generates for its subjects and for society at large? With this framework in mind, consider some examples.

1. Friction Offline

In the physical world speed bumps are perhaps the most intuitive example of friction, generating both a physical element of friction as one drives over them, as well as the behavioral friction involved in deliberately slowing down (or the penalty to be paid when a driver has not done so).⁵⁸ Modern societies have become adept at building smooth roads, allowing riders to drive their cars at high speeds. Given this man-made infrastructure, speed bumps are a physical barrier added to roads to nudge drivers to slow down to promote the human value of mixed traffic on streets — enabling people and cars to safely share the same space. They generate a variety of costs for different actors beyond just the driver. For example, passengers in a car that did not slow down for the speed bump may suffer discomfort. Drivers who wish to avoid slowing down due to the presence of speed bumps may decide to take a different, longer, less convenient route, leaving the road with the speed bumps less traveled by.

Other examples of offline friction include verifying a user's age or identity before allowing them to buy alcohol or to cast a vote, mandated

⁵⁷ Frischmann & Benesch, *supra* note 9, at 19-20.

⁵⁸ *Id.* at 14-19.

disclosure requirements such as detailing the ingredients and nutritional value of food, building codes, and licensing requirements.⁵⁹

2. Friction Online

Our online experience is also dense with different types of friction. Part of what makes frictionless technology so attractive to technology companies is the automatic, instinctive nature by which users are able to complete certain actions. Technology encourages us to respond quickly, to be the first to post content, and to share or retweet immediately. Integrating friction into technology can lead users to change the nature of their response.⁶⁰ Rather than an automatic knee-jerk response, friction encourages users to engage in a more deliberative thought process, leading to a more calculated outcome. Asking users questions like “are you sure you want to delete this file?” or “do you really want to unsubscribe from this mailing list?” encourages users to slow down and exercise cognitive autonomy. It asks users to make a deliberative choice, while at the same time not preordaining the final decision.

One example of integrating friction into technology in order to encourage deliberation involves “take a break” features integrated by leading platforms to fight addiction.⁶¹ Many social media platforms

⁵⁹ *Id.* at 25.

⁶⁰ In this Article we focus on friction in human interface design, particularly on the design and architecture of social media and instant messaging services. Technology companies integrate friction in other aspects of their activity too, having nothing to do with human-computer interaction. See, e.g., Cynthia Dwork & Moni Naor, *Pricing via Processing or Combatting Junk Mail*, in *ADVANCES IN CRYPTOLOGY - CRYPTO '92*, at 139, 139-40 (1992) (proposing the introduction of friction as a tool to identify and screen spam mail by requiring senders' emails to solve a cryptographic puzzle before allowing a message into the receiver's inbox); see also Ohm & Frankle, *supra* note 5, at 791-92 (describing the introduction of friction by the IEX stock exchange, in the form of an eight-mile-long fiber optic, which generated a 350 microsecond delay aimed at eliminating actors seeking to gain an unfair advantage at stock trading).

⁶¹ See Jordan Furlong, *Investing in Our Community's Digital Well-Being*, TIKTOK NEWSROOM (June 9, 2022), <https://newsroom.tiktok.com/en-us/investing-in-our-communitys-digital-well-being> [<https://perma.cc/Z9AP-LYMD>]; Samantha Murphy Kelly, *Instagram Will Now Tell Users When to Take a Break from Using the App*, CNN BUS. (Dec. 7, 2021, 3:00 AM EST), <https://www.cnn.com/2021/12/07/tech/instagram-take-a-break/index.html> [<https://perma.cc/B2QK-T9R3>].

purposely design users' newsfeeds to be seamless and never-ending, encouraging endless scrolling and contributing to social media addiction.⁶² Introducing a structured stop in this endless scrolling has the expected advantage of calling users' attention to their own behavior and encouraging them to put their phone down. Recently TikTok announced that it would introduce "scrolling breaks" for its users, taking part in the fight against social media and cell phone addiction.⁶³ This artificial seam⁶⁴ calls upon users to consciously decide whether they want to be spending *even more* time scrolling through the app. In 2021, Instagram introduced its "Take a Break" tool in response to criticism that the platform was addictive to teens.⁶⁵ This type of friction reminds teens who have spent a long period of time scrolling through the app to take a break. Note that these breaks are suggestions and not mandatory; this is a form of friction that teens can choose to ignore. This form of friction is one of the provisions in the proposed Social Media Addiction Reduction Technology ("SMART") Act.⁶⁶ Section 3(2) of the Act would require platforms to create such a seam in their newsfeed after content that "the typical user scrolls through in 3 minutes."⁶⁷

Leading social media platforms use friction in other contexts as well. In May 2020, Twitter started testing a new tool to warn users before

⁶² See, e.g., Arvind Narayanan, *TikTok's Secret Sauce*, KNIGHT FIRST AMEND. INST. AT COLUM. UNIV. (Dec. 15, 2022), <https://knightcolumbia.org/blog/tiktoks-secret-sauce> [<https://perma.cc/BAV4-4U4D>] (describing the way TikTok's algorithm encourages ongoing viewing, eliminating the need to click on videos).

⁶³ Stephanie Mlot, *TikTok Starts Reminding Users to Stop Scrolling and Go Do Something Else*, PCMAG (June 9, 2022), <https://www.pcmag.com/news/tiktok-starts-reminding-users-to-stop-scrolling-and-go-do-something-else> [<https://perma.cc/PX4C-CT94>].

⁶⁴ Frischmann & Ohm, *supra* note 9 (manuscript at 1).

⁶⁵ Adam Mosseri, *Raising the Standard for Protecting Teens and Supporting Parents Online*, INSTAGRAM (Dec. 7, 2021), <https://about.instagram.com/blog/announcements/raising-the-standard-for-protecting-teens-and-supporting-parents-online> [<https://perma.cc/YVY2-KPTJ>].

⁶⁶ Social Media Addiction Reduction Technology (SMART) Act, S. 2314, 116th Cong. (2019).

⁶⁷ *Id.* § 3(2).

they replied to a tweet with language that was considered “harmful.”⁶⁸ The tool was adopted and implemented across all users a year later.⁶⁹ During that year, Twitter reported finding that thirty-four percent of people presented with this prompt edited their response or decided not to post it at all.⁷⁰ One of the challenges the system had to contend with was the nuance of human conversations. Words that in one context may seem offensive, may not be in a different context. To overcome this, Twitter looked not only at the words being used but also at the relationship between the original poster and the responding one.⁷¹ One criterion the system took into consideration when deciding whether to flag language as offensive was the frequency of past correspondence between the two parties. A word used casually in the context of an ongoing relationship may be seen as inappropriate outside of a relationship.⁷² Like Instagram’s “Take a Break,” there is no ban on using certain words and no technological block preventing a user from using harmful language.⁷³ This prompt is a modern-day form of friction many of us were taught as children — to count to ten before we say something unkind.⁷⁴

In June 2020, Twitter experimented with yet another new type of friction in the form of a prompt. Before a user retweeted an article without first clicking on the link, they were presented with a prompt

⁶⁸ Twitter Support (@TwitterSupport), TWITTER (May 5, 2020, 1:01 PM EST), <https://twitter.com/TwitterSupport/status/1257717113705414658> [<https://perma.cc/56DG-D6XD>].

⁶⁹ Anita Butler & Alberto Parrella, *Tweeting with Consideration*, TWITTER BLOG (May 5, 2021), https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration [<https://perma.cc/3ZMT-5NMG>].

⁷⁰ *Id.*

⁷¹ *Id.*

⁷² See Sam Machkovech, *Twitter’s Latest Robo-Nag Will Flag “Harmful” Language Before You Post*, ARS TECHNICA (May 5, 2021, 4:10 PM), <https://arstechnica.com/information-technology/2021/05/twitters-latest-robot-nag-will-flag-harmful-language-before-you-post/> [<https://perma.cc/P5BP-U8K9>].

⁷³ See Kelly, *supra* note 61.

⁷⁴ The following quote is attributed to Thomas Jefferson: “When angry, count to ten before you speak. If very angry – a hundred.”

asking if they would like to open the article before retweeting it.⁷⁵ Twitter's stated goal with this tool was to "promote informed discussion."⁷⁶ The reported results were encouraging: people click on the link forty percent more often when presented with the prompt.⁷⁷ In addition, some users change their mind about retweeting an article once they opened it, though Twitter did not report the exact percentage of users who made this decision.⁷⁸ It is also unclear whether users decided to refrain from retweeting because they opened the article, read it, found that the headline did not indeed reflect the content of the article and then decided it was not worthy of retweeting, or whether the friction alone discouraged the retweeting.

The friction in these cases serves as "nudges," the term for forms of soft paternalism by which policymakers (technology companies in this case) try to encourage users to behave in a certain way.⁷⁹ Nudges do not prevent users from making poor choices. Instead, they make it more difficult and cumbersome to select the choice that the policymaker disfavors.⁸⁰

These examples and many more amount to a new trend in which technology companies are integrating friction into core features of their design. Though friction is not without precedent — software products have deployed limited amounts of friction to discourage spam and poor

⁷⁵ Twitter Support (@TwitterSupport), TWITTER (June 10, 2020, 2:23 PM EST), <https://twitter.com/TwitterSupport/status/1270783537667551233> [<https://perma.cc/N3NY-3QQK>].

⁷⁶ *Id.*

⁷⁷ Twitter Support (@TwitterSupport), TWITTER (Sept. 24, 2020, 1:11 PM EST), <https://twitter.com/TwitterSupport/status/1309178716988354561> [<https://perma.cc/2NBD-P54G>].

⁷⁸ *Id.*

⁷⁹ See RICHARD H. THALER & CASS R. SUNSTEIN, *NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS* 47 (2008); EYAL ZAMIR & BARAK MEDINA, *LAW, ECONOMICS AND MORALITY* 316-17 (2010); Lauren E. Willis, *When Nudges Fail: Slippery Defaults*, 80 U. CHI. L.REV. 1155, 1158 (2013). A paternalistic policy maker believes they know better than the subjects of the relevant policy what the desired course of action for them is. Jeffrey Rachlinski, *The Uncertain Psychological Case for Paternalism*, 97 NW. U. L. REV. 1165, 1166 (2003).

⁸⁰ See THALER & SUNSTEIN, *supra* note 79, at 8.

cybersecurity practices⁸¹ — the introduction of friction into the core operation of communications platforms is a break away from the frictionless tendencies technology companies have heralded for many years. These questions have not received sufficient systemic scholarly or policy attention to date. Inasmuch as regulators may want to start mandating friction as a tool to overcome challenges created by technology, it is imperative that such questions be asked, studied, and answered. To start, we study the growing use of friction as an important tool for dealing with one of the greatest challenges facing society today, the uncontrolled spread of false news.

II. FALSE NEWS

Social media and instant messaging services, which have developed rapidly in recent years, generate various harms to their users and to society at large.⁸² Amongst all these harms (including manipulation, polarization, extremism) some believe that false news is “*the* defining political communication topic of our time.”⁸³ In this article, we use the term “false news” as used by Vosoughi, Roy, and Aral.⁸⁴ We mean for this term to encompass the three main types of false news discussed in the literature, namely: misinformation, disinformation, and fake news.⁸⁵ In this paper we adopt a broad interpretation of the term “false news” to

⁸¹ See, e.g., Elie Bursztein, Matthieu Martin & John C. Mitchell, *Text-Based CAPTCHA Strengths and Weaknesses*, in PROCEEDINGS OF THE 18TH ACM CONFERENCE ON COMPUTER & COMMUNICATIONS SECURITY 125 (2011) (describing the use of CAPTCHAs to block automated interactions with websites); Verena Distler, Gabriele Lenzini, Carine Lallemand & Vincent Koenig, *The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely*, in NEW SECURITY PARADIGMS WORKSHOP 2020, at 45, 46-47 (2020) (discussing “nudges” and other types of friction).

⁸² See Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown & Alexander Volfovsky, *Exposure to Opposing Views on Social Media Can Increase Political Polarization*, 37 PROC. NAT’L ACAD. SCI. 9216, 9217 (2018); Gordon-Tapiero et al., *supra* note 28, *passim*; Daniel Susser, Beate Roessler & Helen Nissenbaum, *Technology, Autonomy, and Manipulation*, 8 INTERNET POL’Y REV. 1, 8-11 (2019); Zeynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 COLO. TECH. L.J. 203 *passim* (2015).

⁸³ Freelon & Wells, *supra* note 2, at 145.

⁸⁴ Soroush Vosoughi, Deb Roy & Sinan Aral, *The Spread of True and False News Online*, 359 SCIENCE 1146 (2018).

⁸⁵ See discussion *infra* Part II.A.2 on terminology.

describe content that is manipulative, false or in any other way untrue. While this includes news stories, it is by no means limited to them. We also recognize that the effectiveness of friction as a tool to fight false news may be different when applied to the different types of false news (i.e. friction is likely to be more effective in curbing the spread of *misinformation* compared to cases of *disinformation*, spread by motivated users).⁸⁶

In this part we review the phenomenon of false news, the harms it has caused, and the psychological sources driving it.

A. *The Harms of False News*

The publication and spread of false news is far from being a modern-day invention. In August 1835, readers of *The Sun* newspaper were shocked to learn from an article that life had been discovered on the moon!⁸⁷ Five more articles followed detailing the astonishing discovery: unicorns, two legged beavers, and human-like creatures with bat wings all existed and lived on the moon.⁸⁸ Several weeks after the publication, *The Sun* admitted the articles were a hoax and that no such discoveries had been made.⁸⁹ In the decades since then, countless other false reports have been made. Some have appeared in mainstream media sources, while others have circled between friends and communities, spreading as rumors via word of mouth.⁹⁰ With the development of technology, social media, and instant messaging apps, false news has taken on a different nature, a changed pattern of circulation as well as an increased

⁸⁶ See discussion *infra* Parts V.C.2, V.D.2.

⁸⁷ See István Kornél Vida, *The “Great Moon Hoax” of 1835*, 18 HUNGARIAN J. ENG. & AM. STUD. 431, 431 (2012); see also Pennycook & Rand, *supra* note 12, at 388.

⁸⁸ “*The Great Moon Hoax*” Is Published in the “*New York Sun*,” HISTORY: THIS DAY IN HISTORY (Nov. 24, 2009), <https://www.history.com/this-day-in-history/the-great-moon-hoax> [<https://perma.cc/K3PC-A38N>].

⁸⁹ *Id.*

⁹⁰ See CRAIG SILVERMAN, COLUM. JOURNALISM R.: TOW CTR. FOR DIGIT. JOURNALISM, LIES, DAMN LIES, AND VIRAL CONTENT 11, 28 (2015), https://www.cjr.org/tow_center_reports/craig_silverman_lies_damn_lies_viral_content.php [<https://perma.cc/VB8V-YDPS>]; see also Catherine Beauvais, *Fake News: Why Do We Believe It?*, 89 JOINT BONE SPINE 1, 2 (2022).

speed, reach, and threat.⁹¹ Platforms' ability to target false content to those users who are most likely to be susceptible to believing it, makes it a larger concern now than it was in the past.⁹²

1. The Societal Nature of the Harms

False news not only harms individuals:⁹³ it impacts the very social and political fabric of our communities.⁹⁴ The negative impact of false news on society manifests itself in several types of harms: negatively impacting public discourse and deliberation processes, undermining individuals' ability to formulate a joint sense of reality or truth, and ultimately chipping away at the public's belief in democracy and its underlying processes and institutions.⁹⁵ Today's increasing political polarization is not only the result of false news but also fertile ground for the spread of belief in sensationalist, extreme, and polarizing stories.⁹⁶

⁹¹ See Lili Levi, *Real "Fake News" and Fake "Fake News,"* 16 FIRST AMEND. L. REV. 232, 236 (2018).

⁹² See *id.* at 253.

⁹³ See An Nguyen & Daniel Catalan-Matamoros, *Digital Mis/Disinformation and Public Engagement with Health and Science Controversies: Fresh Perspectives from Covid-19*, 8 MEDIA & COMM'N 323, 323-24 (2020); CTR. FOR COUNTERING DIGIT. HATE, THE DISINFORMATION DOZEN: WHY PLATFORMS MUST ACT ON TWELVE LEADING ONLINE ANTI-VAXXERS (2021), <https://counterhate.com/wp-content/uploads/2022/05/210324-The-Disinformation-Dozen.pdf> [<https://perma.cc/2EWB-QBPC>]. See generally Felix Simon, Philip N. Howard & Rasmus Kleis Nielsen, *Types, Sources, and Claims of COVID-19 Misinformation*, REUTERS INST. FOR THE STUDY OF JOURNALISM (Apr. 7, 2020), <https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation> [<https://perma.cc/235E-3PUX>] (presenting findings on the prevalence of Covid-19 misinformation).

⁹⁴ See Tomer Shadmy, *Content Traffic Regulation: A Democratic Framework for Addressing Misinformation*, 63 JURIMETRICS J. 1, 10 (2022).

⁹⁵ See Julie E. Cohen, *Tailoring Election Regulation: The Platform is the Frame*, 4 GEO. L. TECH. REV. 641, 659 (2020) (suggesting that the way that public discourse occurs on platforms undermines the structure necessary for a stable democracy); Terry Lee, *The Global Rise of "Fake News" and the Threat to Democratic Elections in the USA*, 22 PUB. ADMIN. & POL'Y: AN ASIA-PAC. J. 15, 20 (2019).

⁹⁶ Lazer et al., *supra* note 8, at 1096; see Levi, *supra* note 91, at 236. See generally Christine Hagar, *Crisis Informatics: Perspectives of Trust – Is Social Media a Mixed Blessing?*, 2 SCH. INFO. STUDENT RSCH. J. 1 (2013) (discussing the potential for social media to spread misinformation).

Forcing people to constantly mistrust content they see can impact and distort the idea of truth. Science and other evidence-based institutions rely on the shared understanding of what is real and what is not, what is reliable and what is not.⁹⁷ Causing individuals to constantly question the content they are faced with may make them question whether a joint objective truth even exists and distrustful of those who set out to find it.⁹⁸

On a community level, the ability to generate a shared sense of reality is a crucial element of any democratic state.⁹⁹ Without a joint sense of reality, many elements of democracy are undermined. First, meaningful debate and public deliberation rests upon a shared notion at least of what is real and what is false. Hannah Arendt warns that loss of a joint sense of reality undermines the very cornerstone of a democratic regime. “Totalitarian propaganda thrives on this escape from reality into fiction [Disinformation] can outrageously insult common sense only where common sense has lost its validity.”¹⁰⁰ The widespread dissemination of false news has resulted in the collective undermining of trust in a reliable and free press, democracy as a legitimate form of government, and of democratic institutions.¹⁰¹ Democracies are not only

⁹⁷ Courts are an example of an important democratic institution that is based on the belief that there is a truth, and it can be reached. Ronnell Andersen Jones & Lisa Grow Sun, *Freedom of the Press in Post-Truthism America*, 98 WASH. U. L. REV. 419, 420 (2020); Levi, *supra* note 91, at 267; *see also* Allison Orr Larsen, *Constitutional Law in an Age of Alternative Facts*, 93 N.Y.U. L. REV. 175, 181-82 (2018).

⁹⁸ Spencer McKay and Chris Tenove suggest calling this effect “epistemic cynicism.” Spencer McKay & Chris Tenove, *Disinformation as a Threat to Deliberative Democracy*, 74 POL. RSCH. Q. 703, 704 (2021); *see also* Shadmy, *supra* note 94, at 11.

⁹⁹ On the impact of disinformation on deliberative democratic processes, *see* McKay & Tenove, *supra* note 98, at 704. On the epistemic function of deliberative systems, which “is to produce preferences, opinions, and decisions that are appropriately informed by facts and logic and are the outcome of substantive and meaningful consideration of relevant reasons,” *see* Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson & Mark E. Warren, *A Systemic Approach to Deliberative Democracy*, in *DELIBERATIVE SYSTEMS: DELIBERATIVE DEMOCRACY AT THE LARGE SCALE* 1, 11 (John Parkinson & Jane Mansbridge eds., 2012).

¹⁰⁰ HANNAH ARENDT, *THE ORIGINS OF TOTALITARIANISM* 352 (1979).

¹⁰¹ *See* Lazer et al., *supra* note 8, at 1094-95; Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles & David G. Rand, *Shifting Attention to Accuracy Can Reduce Misinformation Online*, 592 NATURE 590, 590 (2021); Pennycook & Rand, *supra*

more sensitive to the dangers stemming from the spread of false news, they are also less equipped to restrict it.¹⁰² In particular, false news has the potential to influence individuals' political beliefs, impact public opinion, and ultimately even skew election results.¹⁰³

2016 can be seen as a meaningful point in time regarding the use, prevalence and impact of false news on democratic elections. False news was a central point of concern during the U.S. presidential election that year.¹⁰⁴ Over 125 million Americans were exposed through Facebook and other social networks to false news reports fabricated by the Russian government in an attempt to intervene in the election process.¹⁰⁵ Moreover, Russia operated thousands of fake profiles on social media sites for purposes of “sowing discord in the U.S. political system” and

note 12, at 396; Kate Starbird, Arif Ahmer & Tom Wilson, *Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations*, 3 *PROC. ACM ON HUM.-COMPUT. INTERACTION* 1, 17 (2019); *see also* Loni Hagen, Stephen Neely, Thomas E. Keller, Ryan Scharf & Fatima Espinoza Vasquez, *Rise of the Machines? Examining the Influence of Social Bots on a Political Discussion Network*, 40 *SOC. SCI. COMP. REV.* 264, 265 (2022) (“Of particular concern in recent years has been the growing influence of *social bots* in political discussion networks, notably their potential to adversely impact democratic outcomes.”); P.M. Krafft & Joan Donovan, *Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign*, 37 *POL. COMMUN.* 194, 208 (2020) (discussing the theoretical implications of disinformation).

¹⁰² *See* Shadmy, *supra* note 94, at 11.

¹⁰³ *See* Anthony J. Gaughan, *Illiberal Democracy: The Toxic Mix of Fake News, Hyperpolarization, and Partisan Election Administration*, 12 *DUKE J. CONST. L. & PUB. POL'Y* 57, 59 (2017); Levi, *supra* note 91, at 233, 240 (noting the threat that false news poses to elections); *see also* Larsen, *supra* note 97, at 180 (identifying “new forces at work that should make us concerned that the same disease plaguing today’s political dialogue will infect (or further infect) the judiciary”).

¹⁰⁴ Benjy Sarlin, “Fake News” Went Viral in 2016. *This Expert Studied Who Clicked.*, NBC NEWS (Jan. 14, 2018, 7:06 AM EST), <https://www.nbcnews.com/politics/politics-news/fake-news-went-viral-2016-expert-studied-who-clicked-n836581> [<https://perma.cc/A2NR-UYC4>]. The term “post-truth” was even chosen as Oxford Dictionary’s word of the year for 2016. Amy B. Wang, ‘Post-Truth’ Named 2016 Word of the Year by Oxford Dictionaries, WASH. POST (Nov. 16, 2016, 9:16 AM EST), <https://www.washingtonpost.com/news/the-fix/wp/2016/11/16/post-truth-named-2016-word-of-the-year-by-oxford-dictionaries/> [<https://perma.cc/4BVE-AHM9>].

¹⁰⁵ Carol E. Lee & Jo Ling Kent, *Facebook Says Russian-Backed Election Content Reached 126 Million Americans*, NBC NEWS (Oct. 30, 2017, 3:00 PM PST), <https://www.nbcnews.com/news/us-news/195psos195n-backed-election-content-reached-126-million-americans-facebook-says-n815791> [<https://perma.cc/8TAK-N83X>].

attempting to influence public opinion.¹⁰⁶ As described by Lili Levi, the term “fake news” itself became “the central inflammatory charge in media discourse in the United States since the 2016 presidential contest.”¹⁰⁷ Since then, false news has been a concern in almost every central election around the world.¹⁰⁸

2. Terminology

In this article, following others, we use the term “false news” to encompass several related concepts.¹⁰⁹

¹⁰⁶ ROBERT S. MUELLER III, I REPORT ON THE INVESTIGATION INTO RUSSIAN INTERFERENCE IN THE 2016 PRESIDENTIAL ELECTION 14 (2019) (“The IRA [Internet Research Agency] conducted social media operations targeted at large U.S. audiences with the goal of sowing discord in the U.S. political system.”).

¹⁰⁷ Levi, *supra* note 91, at 233.

¹⁰⁸ See Christoph Bluth, *Why Public Trust in Elections is Being Undermined by Global Disinformation Campaigns*, THE CONVERSATION (Apr. 28, 2022, 10:27 AM EST), <https://theconversation.com/why-public-trust-in-elections-is-being-undermined-by-global-disinformation-campaigns-181825> [<https://perma.cc/D3F6-GU8L>]; Daniel Funke & Daniela Flamini, *A Guide to Anti-Misinformation Actions Around the World*, POYNTER, <https://www.poynter.org/ifcn/anti-misinformation-actions/> (last updated Apr. 9, 2018) [<https://perma.cc/L5CZ-UEKV>]; see also Samantha Lai, *Data Misuse and Disinformation: Technology and the 2022 Elections*, BROOKINGS (June 21, 2022), <https://www.brookings.edu/blog/techtank/2022/06/21/data-misuse-and-disinformation-technology-and-the-2022-elections/> [<https://perma.cc/H743-FCAX>].

¹⁰⁹ See, e.g., Kai Shu, Suhang Wang, Dongwon Lee & Huan Liu, *Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements*, in DISINFORMATION, MISINFORMATION, AND FAKE NEWS IN SOCIAL MEDIA 1, 1 (Kai Shu, Suhang Wang, Dongwon Lee & Huan Liu eds., 2020) (“We take fake news as an example of disinformation. The extensive spread of fake news can have severe negative impacts on individuals and society.”); Ryan Calo, Chris Coward, Emma S. Spiro, Kate Starbird & Jevin D. West, *How Do You Solve a Problem like Misinformation?*, 7 SCI. ADVANCES, Dec. 8, 2021, at 1 (“Disinformation refers to a purposive strategy to induce false belief, channel behavior, or damage trust.”); Levi, *supra* note 91, at 245 (“[T]he phrase is an umbrella term referring to ‘real threats to meaningful public debate on the Internet.’”); Dawn Carla Nunziato, *Misinformation Mayhem: Social Media Platforms’ Efforts to Combat Medical and Political Misinformation*, 19 FIRST AMEND. L. REV. 32, 89-98 (2020) (discussing the consistency of social media platforms’ measures to combat false news with the First Amendment); Mark Verstraete, Jane R. Bambauer & Derek E. Bambauer, *Identifying and Countering Fake News*, 73 HASTINGS L.J. 821, 823 (2022) (“[T]he term has been used to refer to so many things that it seems to have completely lost its power to describe; as a result, several media critics have recommended abandoning the moniker entirely.”);

Disinformation is the term used to describe “purposive strategy to induce false belief, channel behavior, or damage trust.”¹¹⁰ The spread of disinformation is often part of an organized campaign. While there can be many goals for spreading disinformation, empirical research has shown that orchestrated campaigns of disinformation often target election outcomes as well as “undermin[e] the institutions and social conditions necessary for democracies to function.”¹¹¹ To lend increased credibility to disinformation campaigns, their organizers may sometimes make a point of integrating true information into them.¹¹²

The term *misinformation* describes wrong, fake, or misleading content that individuals are exposed to and may erroneously share or engage with, not knowing that the content is false.¹¹³ The outcomes of spreading misinformation can be as harmful as those of the spread of disinformation and include a dire impact on “civic knowledge and . . . the ability to conduct public debates and deliberation processes.”¹¹⁴ One may unintentionally find themselves spreading misinformation, if they are not aware of its falsity or were misled by the content themselves.

Fabio Giglietto, Laura Iannelli, Luca Rossi & Augusto Valeriani, *Fakes, News and the Election: A New Taxonomy for the Study of Misleading Information Within the Hybrid Media System* 30 (Dec. 2, 2016) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2878774 [<https://perma.cc/DKA7-CQSF>] (describing process-centered rather than actor-centered approach to the dissemination of false information online).

¹¹⁰ Calo et al., *supra* note 109, at 1.

¹¹¹ McKay & Tenove, *supra* note 98, at 703; *see also* Krafft & Donovan, *supra* note 101, at 194 (“Disinformation campaigns such as those perpetrated by far-right groups in the United States seek to erode democratic social institutions.”).

¹¹² Calo et al., *supra* note 109, at 1.

¹¹³ *Id.*

¹¹⁴ Shadmy, *supra* note 94, at 7; *see also* Stephan Lewandowsky, Werner G. K. Stritzke, Alexandra M. Freund, Klaus Oberauer & Joachim I. Krueger, *Misinformation, Disinformation, and Violent Conflict: From Iraq and the “War on Terror” to Future Threats to Peace*, 68 AM. PSYCH. 487, 491 (2013) (discussing the link between belief in misinformation and support for the Iraq War); Tomer Simon, Avishay Goldberg, Dmitry Leykin & Bruria Adini, *Kidnapping WhatsApp – Rumors During the Search and Rescue Operation of Three Kidnapped Youth*, 64 COMPUTS. HUM. BEHAV. 183, 183 (2016) (“Social media are used during emergencies to distribute relevant, critical information to the public and the authorities, and may be simultaneously used to distribute rumors, misinformation and unverified data, which propagate rapidly.”).

Fake news is a particular type of disinformation. The term refers to “news articles with intentionally false information, [which] are produced online for a variety of purposes, ranging from financial to political gains.”¹¹⁵ In recent years, the term has increasingly been used by politicians to instead discredit their opponents as unreliable.¹¹⁶

B. *The Far and Wide Reach of False News*

There is widespread agreement among the American public that false news is a serious concern. A survey by Pew Research found that eighty-eight percent of Americans believed that false news is responsible for sowing confusion among the American public.¹¹⁷ At the same time, fewer than twenty-five percent admitted to sharing content they knew was untrue.¹¹⁸ In this section we show how these two findings can be explained.

In Parts III and IV we present and analyze WhatsApp’s integration of friction as a tool to limit the spread of false news. When designing a tool aimed at impacting people’s behavior, it is important to first understand the considerations driving that behavior. Therefore, in this part we present the psychological reasons driving people to believe and to share false news. Believing false news does not necessarily give rise to sharing it, nor does sharing false news hinge on believing it. Ultimately, the friction-creating tools chosen by WhatsApp reflect an attempt to make use of these psychological mechanisms, while still providing users with the ability to make their own choice whether they want to share content or not.

1. Who Believes False News?

How likely are you to be able to identify a fake headline when presented with one? If you are like three in four Americans, you may be

¹¹⁵ Shu et al., *supra* note 109, at 2.

¹¹⁶ See WARDLE & DERAKHSHAN, *supra* note 11, at 5; Vosoughi et al., *supra* note 84, at 1146.

¹¹⁷ Michael Barthel, Amy Mitchell & Jesse Holcomb, *Many Americans Believe Fake News Is Sowing Confusion*, PEW RSCH. CTR. 1, 5 (Dec. 15, 2016), <https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/> [https://perma.cc/8KG8-37PQ].

¹¹⁸ *Id.* at 1.

overestimating your ability to identify false news,¹¹⁹ and if you are like ninety percent of Americans, you believe that your ability to identify false news is above average.¹²⁰ It is widely agreed that false news is prevalent, especially on social media and instant messaging services.¹²¹ If people mistakenly believe that they are able to identify false news, it may make them more susceptible to believing and sharing it.¹²²

Certain characteristics may cause people to be more susceptible to believing false news. First, overconfident individuals are more likely to believe false news,¹²³ while people who are more reflective are less likely to do so.¹²⁴ A second criterion is a message's level of familiarity. The *familiarity bias* describes a cognitive tendency to believe messages that one has been exposed to in the past, more than unfamiliar ones.¹²⁵ Repeatedly exposing individuals to the same false news is likely to increase their belief in it, even if they were not inclined to believe it upon

¹¹⁹ See Benjamin A. Lyons, Jacob M. Montgomery, Andrew M. Guess, Brendan Nyhan & Jason Reifler, *Overconfidence in News Judgments is Associated with False News Susceptibility*, 118 PROC. NAT'L ACAD. SCI., 2021, at 1.

¹²⁰ See *id.*; cf. *More Than 1 in 3 Americans Believe a "Deep State" Is Working to Undermine Trump*, IPSOS (Dec. 30, 2020), <https://www.ipsos.com/en-us/news-polls/npr-misinformation-123020> [<https://perma.cc/V79A-Q3PB>] (“[F]ewer than half (47%) are able to correctly identify that this statement is false: ‘A group of Satan-worshipping elites who run a child sex ring are trying to control our politics and media.’ Thirty-seven percent are unsure whether this theory backed by QAnon is true or false, and 17% believe it to be true.”).

¹²¹ Gizem Ceylan, Ian A. Anderson & Wendy Wood, *Sharing of Misinformation is Habitual, Not Just Lazy or Biased*, 120 PROC. NAT'L ACAD. SCI., 2023, at 1; see SILVERMAN, *supra* note 90, at 12-13.

¹²² See Lyons et al., *supra* note 119, at 1.

¹²³ See *id.*; Nikita A. Salovich, Amalia M. Donovan, Scott R. Hinze & David N. Rapp, *Can Confidence Help Account for and Redress the Effects of Reading Inaccurate Information?*, 49 MEMORY & COGNITION 293, 306 (2021).

¹²⁴ Mohsen Mosleh, Gordon Pennycook, Antonio A. Arechar & David G. Rand, *Cognitive Reflection Correlates with Behavior on Twitter*, 12 NATURE COMM'NS, 2021, at 1; Pennycook & Rand, *supra* note 12, at 392; see Jordan Carpenter, Daniel Preotiuc-Pietro, Jenna Clark, Lucie Flekova, Laura Smith, Margaret L. Kern, Anneke Buffone, Lyle Ungar & Martin Seligman, *The Impact of Actively Open-Minded Thinking on Social Media Communication*, 13 JUDGMENT DECISION MAKING 562, 571 (2018).

¹²⁵ Ullrich K.H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga & Michelle A. Amazeen, *The Psychological Drivers of Misinformation Belief and its Resistance to Correction*, 1 NATURE REV. PSYCH. 13, 14 (2022).

the first exposure.¹²⁶ According to this theory, then, broadly circulated content that a user is exposed to more than once, say on social media or instant messaging, would have an increased chance of generating belief. Third, the source of news has a large impact on its believability. Content provided by someone deemed as credible is more likely to be believed.¹²⁷ Fourth, headlines that evoke a strong emotional reaction (shock, fear, anger, outrage) are more likely to engender belief.¹²⁸ Finally, the content of the news also impacts people's susceptibility to believing it. Overconfident individuals are more likely to believe false political news if it aligns with their pre-existing beliefs,¹²⁹ while the ability of reflective individuals to discern between false and true news is not impacted by their previously held political positions.¹³⁰ Researchers also found that increased deliberation reduced users' belief in false news.¹³¹

In sum, careful reasoning and deliberation increase individuals' ability to distinguish between false and true news. Lack of knowledge, a strong emotional response or the familiarity heuristic all have a negative impact on the ability to identify false news. Based on an analysis of all these barriers to identifying false news, Gordon Pennycook and David Rand suggest that the right type of intervention can encourage social media users to focus their attention more on evaluating the accuracy of the content they consume and share, thus lowering the prevalence of false news.¹³²

¹²⁶ See Gordon Pennycook & David G. Rand, *Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking*, 88 J. PERSONALITY 185, 186 (2020); Pennycook & Rand, *supra* note 12, at 393.

¹²⁷ Pennycook & Rand, *supra* note 12, at 393.

¹²⁸ *Id.*; see Anastasia Kozyreva, Stephan Lewandowsky & Ralph Hertwig, *Citizens Versus the Internet: Confronting Digital Challenges with Cognitive Tools*, 21 PSYCH. SCI. PUB. INT. 103, 124 (2020); Thorsten Quandt, *Dark Participation*, 6 MEDIA COMM'N 36, 42 (2018).

¹²⁹ See Lazer et al., *supra* note 8, at 1095; Lyons et al., *supra* note 119, at 2.

¹³⁰ Pennycook & Rand, *supra* note 12, at 392.

¹³¹ *Id.*; see Bence Bago, David G. Rand & Gordon Pennycook, *Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines*, 149 J. EXPERIMENTAL PSYCH. GEN. 1608, 1611 (2020).

¹³² See Pennycook & Rand, *supra* note 12, at 396-99.

2. Who Spreads False News?

The above analysis pertains to users on an individual level. One of the reasons that false news is such a serious concern on a societal level, however, is not only that it somehow manages to reach certain individuals, but rather that it spreads like an infectant virus. Research conducted by MIT scholars, analyzing over 126,000 tweets on Twitter, found that false news spreads faster and farther than true news.¹³³

The initial dissemination of disinformation is done by a motivated actor, such as a foreign government, a political lobby, or a group seeking to otherwise promote their political or economic interests.¹³⁴ The effectiveness of disinformation largely rests upon its ability to continue being spread by unsuspecting individuals, perpetuating the virality of the content. Some people may share false news because they mistakenly believe it is true,¹³⁵ but the mere fact that someone has shared content does not mean they believe it to be true.¹³⁶ In fact, some people do not even consider the question of accuracy before sharing content online.¹³⁷ As one WhatsApp user in India described: “if you are forwarding messages, especially if it’s an opinion or a rumour, it spreads very quickly. . . . *Sometimes they don’t even read the entire message and just forward it.*”¹³⁸ It is not just that people aren’t very good at differentiating between false and true news. It is that many do not give enough thought

¹³³ Vosoughi et al., *supra* note 84, at 1146.

¹³⁴ Shu et al., *supra* note 109, at 2.

¹³⁵ In such a case, the content would be considered *misinformation*, regardless of its initial source. See Tom Chatfield, *Why We Believe Fake News*, BBC FUTURE (Sept. 8, 2019), <https://www.bbc.com/future/article/20190905-how-our-brains-get-overloaded-by-the-21st-century> [<https://perma.cc/R2TC-XG24>].

¹³⁶ See Pennycook et al., *supra* note 101, at 590.

¹³⁷ See Pennycook & Rand, *supra* note 12, at 395.

¹³⁸ SHAKUNTALA BANAJI, RAM BHAT, ANUSHI AGARWAL, NIHAL PASSANHA & MUKTI SADHANA PRAVIN, WHATSAPP VIGILANTES: AN EXPLORATION OF CITIZEN RECEPTION AND CIRCULATION OF WHATSAPP MISINFORMATION LINKED TO MOB VIOLENCE IN INDIA 27 (2019), <https://www.lse.ac.uk/media-and-communications/assets/documents/research/projects/WhatsApp-Misinformation-Report.pdf> [<https://perma.cc/U8VU-XPYE>] (emphasis added).

to the question of whether news is true or false before sharing it.¹³⁹ Sometimes they may not even read past the headline.¹⁴⁰ A recent study found that asking users to rate the accuracy of a headline before sharing it induced a reflective process and lowered the number of people who rated it as fake and were willing to share it.¹⁴¹ Thus, shifting users' attention to the accuracy of the content they are sharing "increases the quality of news that people subsequently share."¹⁴² Many people are extremely concerned about the impact of false news,¹⁴³ some even citing it as one of the biggest threats to democracy.¹⁴⁴ It is clear that a new system of safeguards, protecting democracy from false news, is imperative.¹⁴⁵ One promising source is friction.

III. FRICTION AS A TOOL TO LIMIT THE SPREAD OF FALSE NEWS

We focus on the WhatsApp messaging service, which has over two billion users worldwide and is the central avenue of communication as

¹³⁹ See *id.* at 44 ("[T]he users who want to 'post first' or 'forward first' are less concerned with reliability or the potential to start dangerous rumours than with immediacy and impact.").

¹⁴⁰ *Id.* at 41.

¹⁴¹ Pennycook & Rand, *supra* note 12, at 395; see Pennycook et al., *supra* note 101, at 592; see also Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu & David G. Rand, *Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention*, 31 PSYCH. SCI. 770, 770 (2020) ("[W]e present evidence that people share false claims about COVID-19 partly because they simply fail to think sufficiently about whether or not the content is accurate when deciding what to share.").

¹⁴² Pennycook et al., *supra* note 101, at 590.

¹⁴³ Lyons et al., *supra* note 119, at 1.

¹⁴⁴ See Lee, *supra* note 95, at 15; McKay & Tenove, *supra* note 98 at 703; Merten Reglitz, *Fake News and Democracy*, 22 J. Ethics & Soc. Phil. 162, 162 (2022); Gabriel R. Sanchez & Keesha Middlemass, *Misinformation Is Eroding the Public's Confidence in Democracy*, Brookings Inst. FixGov (July 26, 2022), <https://www.brookings.edu/blog/fixgov/2022/07/26/misinformation-is-eroding-the-publics-confidence-in-democracy/> [<https://perma.cc/L8GZ-79CT>]; see, e.g., João Pedro Baptista & Anabela Gradim, "Brave New World" of Fake News: How It Works, 28 Javnost: Pub.: J. Eur. Inst. Comm'n & Culture 426, 426 (2021) ("Fake news has become a global threat.").

¹⁴⁵ See Lazer et al., *supra* note 8, at 4.

well as a primary source for news consumption in many countries.¹⁴⁶ We describe the backdrop against which WhatsApp's restrictions on forwarding were developed. While these changes have received extensive news coverage, their technical basis in WhatsApp's software has remained largely under scrutinized by the media, researchers, and policymakers.

A. WhatsApp — Overview

WhatsApp is a messaging service which enables its users to send messages and conduct voice and video calls with one another.¹⁴⁷ As of January 2022 WhatsApp had over two billion users worldwide and is ranked as the most popular messenger app in the world.¹⁴⁸ More than 100 billion messages are sent on WhatsApp daily, up from one billion messages per day in October 2011.¹⁴⁹ WhatsApp operates across all major mobile operating systems, making it easier to communicate with contacts from various countries around the world who may use Android, iOS, or another operating system. WhatsApp also has a web browser version, allowing users to operate the app not just from their phone, but from a computer — and any device with a web browser.¹⁵⁰ An individual

¹⁴⁶ *Two Billion Users – Connecting the World Privately*, WHATSAPP BLOG (Feb. 12, 2020), <https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately> [<https://perma.cc/5J7X-RGYA>].

¹⁴⁷ *Get Started*, WHATSAPP HELP CTR., <https://faq.whatsapp.com/497209988909970/> (last visited July 7, 2023) [<https://perma.cc/XB8G-XQBW>].

¹⁴⁸ *Most Popular Global Mobile Messaging Apps as of January 2023*, STATISTA (Aug. 29, 2023), <https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/> [<https://perma.cc/UZU7-D6D6>]; *Number of Unique WhatsApp Mobile Users Worldwide from January 2020 to June 2023*, STATISTA (July 24, 2023), <https://www.statista.com/statistics/1306022/whatsapp-global-unique-users/> [<https://perma.cc/5ZVH-45FJ>].

¹⁴⁹ Drew Olanoff, *WhatsApp Users Now Send Over One Billion Messages a Day*, THE NEXT WEB (Oct. 31, 2011, 7:52 PM), <https://thenextweb.com/news/whatsapp-users-now-send-over-one-billion-messages-a-day> [<https://perma.cc/Q8EA-GDA8>]; Brittany Vincent, *WhatsApp Reaches 100 Billion Daily Message Milestone*, PCMAG (Nov. 1, 2020), <https://www.pcmag.com/news/whatsapp-reaches-100-billion-daily-message-milestone> [<https://perma.cc/529R-QNRY>].

¹⁵⁰ *About WhatsApp Web and Desktop*, WHATSAPP HELP CTR., https://faq.whatsapp.com/668538004658079/?cms_platform=web (last visited July 7, 2023) [<https://perma.cc/HW8L-BLSG>].

corresponding on WhatsApp can communicate with others either by directly messaging a particular user or by sending messages to a group. Groups are limited in size — currently capped at 1,024 participants.¹⁵¹ There is no explicit restriction on the number of groups each user can join and each user is free to leave a group at any point.¹⁵²

Unlike other social media platforms, the communications channel on WhatsApp is encrypted end-to-end.¹⁵³ This type of encryption prevents third parties, including Meta itself, from “having plaintext access to messages or calls.”¹⁵⁴ End-to-end encryption provides a high level of privacy and cybersecurity protection for WhatsApp users and has made the service an attractive communication outlet for people seeking to avoid surveillance. In non-democratic countries, WhatsApp has been used as a tool to encourage free democratic elections, to organize demonstrations, and even to document and report on the horrors of war.¹⁵⁵ It is also a meaningful tool for immigrants to keep in touch with their families as well as for refugees to share tips and warnings with others in their position.¹⁵⁶ It has also been used, however, by people

¹⁵¹ *How to Create and Invite into a Group*, WHATSAPP HELP CTR., https://faq.whatsapp.com/3242937609289432/?cms_platform=web (last visited July 10, 2023) [<https://perma.cc/R62H-SKRR>]. In this article, we use the term “chat” to describe both one-on-one communications between two individuals as well as messages shared in a group setting.

¹⁵² See Melo et al., *supra* note 13, at 377 n.7.

¹⁵³ Marcelo Santos & Antoine Faure, *Affordance Is Power: Contradictions Between Communicational and Technical Dimensions of WhatsApp’s End-to-End Encryption*, SOC. MEDIA + SOC’Y, July–Sept. 2018, at 6.

¹⁵⁴ WHATSAPP, WHATSAPP ENCRYPTION OVERVIEW 3 (2023), <https://www.whatsapp.com/security/WhatsApp-Security-Whitepaper.pdf> [<https://perma.cc/9L75-XEB3>].

¹⁵⁵ See Lauren Said-Moorhouse, *How Syrian Activists Use WhatsApp to Tell the World Their Stories*, CNN (Sept. 28, 2016, 4:30 AM EDT), <https://www.cnn.com/2016/09/28/middleeast/whatsapp-syria-aleppo-activists/index.html> [<https://perma.cc/K2WE-4JEG>].

¹⁵⁶ Farhad Manjoo, *For Millions of Immigrants, a Common Language: WhatsApp*, N.Y. TIMES (Dec. 21, 2016), <https://www.nytimes.com/2016/12/21/technology/for-millions-of-immigrants-a-common-language-whatsapp.html> [<https://perma.cc/4N8M-2F8Q>]; see Tamoá Calzadilla, *WhatsApp: A Lifeline for New Venezuelan Immigrants in Miami*, UNIVISION (July 5, 2016, 1:03 PM EDT), <https://www.univision.com/univision-news/immigration/whatsapp-a-lifeline-for-new-venezuelan-immigrants-in-miami> [<https://perma.cc/CV4X-3E8C>]; Hanna Kozłowska, *The Most Crucial Item that Migrants and Refugees Carry Is a Smartphone*, QUARTZ (Sept. 14, 2015), <https://qz.com/500062/the-most-crucial-item-that-migrants-and-refugees-carry-is-a-smartphone> [<https://perma.cc/>

enjoying the ability to remain undetected by authorities, as a powerful tool to spread false news undermining democratic institutions.¹⁵⁷

B. Limiting Forwarding as a Way to Fight False News

1. False News on WhatsApp

A study reported that false news on WhatsApp tended to become more viral than real and reliable information.¹⁵⁸ The authors defined virality according to the number of times messages are shared, the number of users sharing them, and the number of public groups they are shared in.¹⁵⁹ The finding that false news spreads faster and wider than real news is in line with earlier findings by other researchers who reached similar results in an experiment conducted on Twitter.¹⁶⁰ They found that false news was seventy percent more likely to be retweeted than true news.¹⁶¹

In another experiment, researchers tracked over 150,000 images in public groups in Indonesia, Brazil, and India.¹⁶² They found that eighty percent of the images appeared on WhatsApp for two days or less after

8J6E-T4BU]. WhatsApp can lower the costs of activism. Emiliano Treré, *Reclaiming, Proclaiming, and Maintaining Collective Identity in the #YoSoy132 Movement in Mexico: An Examination of Digital Frontstage and Backstage Activism Through Social Media and Instant Messaging Platforms*, 18 INFO., COMMUN & SOC'Y 901, 911 (2015). It can also be used to avoid state surveillance. See Amelia Johns & Niki Cheong, *Feeling the Chill: Bersih 2.0, State Censorship, and "Networked Affect" on Malaysian Social Media 2012–2018*, SOC. MEDIA + SOC'Y, Apr.–June 2019, at 6.

¹⁵⁷ See Farhad Manjoo, *The Problem with Fixing WhatsApp? Human Nature Might Get in the Way*, N.Y. TIMES (Oct. 24, 2018), <https://www.nytimes.com/2018/10/24/technology/fixing-whatsapp-disinformation-human-nature.html> [<https://perma.cc/TCE4-N84M>]; Natalie Pang & Yue Ting Woo, *What About WhatsApp? A Systematic Review of WhatsApp and Its Role in Civic and Political Engagement*, FIRST MONDAY, (Jan. 5, 2020), <https://firstmonday.org/ojs/index.php/fm/article/view/10417/8322> [<https://perma.cc/ZN2Y-DPMW>].

¹⁵⁸ Gustavo Resende, Philippe Melo, Julio C.S. Reis, Marisa Vasconcelos, Jussara M. Almeida & Fabrício Benevenuto, *Analyzing Textual (Mis)Information Shared in WhatsApp Groups*, in PROCEEDINGS OF THE 10TH ACM CONFERENCE ON WEB SCIENCE 225 (2019).

¹⁵⁹ *Id.*

¹⁶⁰ Vosoughi et al., *supra* note 84, at 1146.

¹⁶¹ *Id.*

¹⁶² See Melo et al., *supra* note 13, at 376.

their initial appearance, and sixty percent of those were shared for less than 1,000 minutes after they made their first appearance.¹⁶³ The study's authors concluded that "WhatsApp is a very dynamic network and most of its image content is ephemeral, i.e., the images usually appear and vanish quickly."¹⁶⁴ These findings support the idea that sharing images in WhatsApp groups is largely the product of snap decisions to forward content received, rather than of deep deliberation.

Tragically, several conspiracy theories leading to violence and even death have been spread on WhatsApp. In India, hundreds of millions of people use WhatsApp, and the platform has been used to spread false news leading to horrific violence. In May 2017 seven men in India were beaten to death by crowds who believed them to be child and organ traffickers, due to rumors spread on WhatsApp.¹⁶⁵ Over the first six months of 2018, at least twenty-four people in India were murdered under similar circumstances,¹⁶⁶ and the violence continued into the

¹⁶³ *Id.* at 377.

¹⁶⁴ *Id.*

¹⁶⁵ Gowhar Farooq, *Politics of Fake News: How WhatsApp Became a Potent Propaganda Tool in India*, 9 MEDIA WATCH 106, 106 (2018); Annie Gowen, *As Mob Lynchings Fueled by WhatsApp Messages Sweep India, Authorities Struggle to Combat Fake News*, WASH. POST (July 2, 2018, 4:58 PM EDT), https://www.washingtonpost.com/world/asia_pacific/as-mob-lynchings-fueled-by-whatsapp-sweep-india-authorities-struggle-to-combat-fake-news/2018/07/02/683a1578-7bba-11e8-ac4e-421ef7165923_story.html [<https://perma.cc/3E6D-HM37>]; Annie Gowen & Elizabeth Dvoskin, *WhatsApp Launches New Controls After Widespread App-Fueled Mob Violence in India*, WASH. POST (July 19, 2018, 10:53 PM EDT), https://www.washingtonpost.com/world/whatsapp-launches-new-controls-after-widespread-app-fueled-mob-violence-in-india/2018/07/19/64433ec9-c944-446f-8d82-8498234ee8a9_story.html [<https://perma.cc/WY3L-MYDN>]; Danish Raza, "I Saw It on WhatsApp": Why People Believe Hoaxes on the Messaging App, HINDUSTAN TIMES (June 16, 2017, 9:51 AM IST), <https://www.hindustantimes.com/india-news/i-saw-it-on-whatsapp-why-people-believe-hoaxes-on-the-messaging-app/story-FiRtEOi7UvxpzrzoJ7nnK.html> [<https://perma.cc/977J-N7JM>]; see also Simi Bassi & Joyita Sengupta, *Lynchings Sparked by WhatsApp Child-Kidnap Rumours Sweep Across India*, CBC NEWS (July 8, 2018, 1:00 AM PDT), <https://www.cbc.ca/news/world/india-child-kidnap-abduction-video-rumours-killings-1.4737041> [<https://perma.cc/9X8Y-UQ9B>].

¹⁶⁶ See Vijaita Singh & Yuthika Bhargava, *Nothing but Lies: Fake Videos, Rumour Set off the Lynch Mobs*, HINDU (July 7, 2018, 10:21 PM IST), <https://www.thehindu.com/news/national/nothing-but-lies-fake-videos-rumour-set-off-the-lynch-mobs/article61516458.ece> [<https://perma.cc/84NJ-JVMQ>].

second half of 2018.¹⁶⁷ In one of the cases, the violence was perpetrated by villagers spurred by child abduction rumors on WhatsApp.¹⁶⁸ Some rumors were even reinforced by an Indian politician.¹⁶⁹ In an attempt to alert people of the fake nature of these conspiracy theories, Indian authorities sent individuals armed with loudspeakers to villages. In one case villagers beat these individuals.¹⁷⁰

2. WhatsApp Limits Forwards to Combat False News

In early July 2018, the Indian government called on WhatsApp to assume “accountability and responsibility” for allowing “repeated circulation of such provocative content” leading to the murders.¹⁷¹ On July 19, 2018, WhatsApp announced changes to its forwarding policy.¹⁷² The platform announced that it was “launching a test to limit forwarding that will apply to everyone using WhatsApp.”¹⁷³ It set a

¹⁶⁷ See Manjoo, *supra* note 157; Eli Meixler, *Five Killed in Latest Mob Attack After Rumors on Social Media. Here’s What to Know About India’s WhatsApp Murders*, TIME (July 3, 2018, 4:22 AM EDT), <https://time.com/5329030/india-whatsapp-murders-mob-false-rumors/> [<https://perma.cc/VX75-563B>].

¹⁶⁸ *How WhatsApp Helped Turn an Indian Village into a Lynch Mob*, BBC NEWS (July 19, 2018), <https://www.bbc.com/news/world-asia-india-44856910> [<https://perma.cc/R9RC-968Q>].

¹⁶⁹ Saumen Sarker, *Tripura BJP Minister Ratan Lal Nath Spreading Rumours of Kidney Smuggling Racket*, YOUTUBE (June 29, 2018), <https://www.youtube.com/watch?v=4QPAjbQm-ms> [<https://perma.cc/6NWA-CPZ5>]; *Tripura Minister Defends “International Kidney Racket” Statement*, NE. NOW NEWS (July 1, 2018, 5:05 PM), <https://nenow.in/north-east-news/tripura-minister-defends-international-kidney-racket-statement.html> [<https://perma.cc/5Z2F-YB8N>].

¹⁷⁰ Gowen, *supra* note 165 (“To control the subsequent violence, state authorities hired ‘rumor busters,’ including Sukanta Chakraborty, 33, a musician who was paid about \$8 a day to travel from village to village in a van equipped with a loudspeaker, warning of the dangers of fake news. He and two others were beset by a mob wielding bricks and bamboo sticks in a crowded market Thursday.”).

¹⁷¹ Lucia Binding, *India Asks WhatsApp to Curb Fake News Following Spate of Lynchings*, Sky News (July 4, 2018, 12:53 PM), <https://news.sky.com/story/india-asks-whatsapp-to-curb-fake-news-following-spate-of-lynchings-11425849> [<https://perma.cc/V4F9-ZCY7>].

¹⁷² In a help page, WhatsApp stated that the forwarding restriction was aimed, inter alia, at restricting the spread of false news. *About Forwarding Limits*, *supra* note 19.

¹⁷³ *More Changes to Forwarding*, WHATSAPP BLOG (July 19, 2018), <https://blog.whatsapp.com/more-changes-to-forwarding> [<https://perma.cc/XU43-BYRW>] [hereinafter *More Changes to Forwarding*].

stricter forwarding limitation in India, allowing users to forward content to only five chats at once.¹⁷⁴ It also removed a previously introduced “quick forward” button.¹⁷⁵

During the same time period, Brazil experienced its own challenges with WhatsApp. Ahead of the 2018 presidential election, WhatsApp was used to extensively spread election-related false news.¹⁷⁶ Reports found that supporters of one of the presidential candidates paid over three million USD for an organized disinformation campaign on WhatsApp.¹⁷⁷ A similar pattern would later emerge in the elections in Nigeria and India, both dubbed “WhatsApp elections.”¹⁷⁸ Despite calls from

¹⁷⁴ Pranav Dixit, *WhatsApp Is Putting Limits On Forwards After Rumors Spreading Through Its Platform Incited Violence in India*, BUZZFEED NEWS (July 19, 2018, 10:03 PM), <https://www.buzzfeednews.com/article/pranavdixit/whatsapp-limits-forwarding> [<https://perma.cc/F54B-EF7Q>]; Alex Hern, *WhatsApp to Restrict Message Forwarding After India Mob Lynchings*, GUARDIAN (July 20, 2018, 9:41 AM EDT), <https://www.theguardian.com/technology/2018/jul/20/whatsapp-to-limit-message-forwarding-after-india-mob-lynchings> [<https://perma.cc/Z93S-7LYD>]; Rishi Iyengar, *WhatsApp Is Adding New Restrictions as Killings Continue in India*, CNN BUS. (July 20, 2018, 11:03 AM EST), <https://money.cnn.com/2018/07/20/technology/whatsapp-india-mob-lynching/index.html> [<https://perma.cc/2WWV-VKF7>]. In other countries the limit remained at 20. Alex Hern, *WhatsApp to Impose New Limit on Forwarding to Fight Fake News*, GUARDIAN (Apr. 7, 2020, 3:00 AM EDT), <https://www.theguardian.com/technology/2020/apr/07/whatsapp-to-impose-new-limit-on-forwarding-to-fight-fake-news> [<https://perma.cc/3QJW-223K>].

¹⁷⁵ Aside from limiting forwarding, the Indian government also requested WhatsApp enable traceability of messages to identify their source. WhatsApp rejected this request, explaining that end-to-end encryption, a critical component of the app, made traceability impossible. T.N.N., *WhatsApp Agrees to Meet all Govt Demands Except Message Traceability*, TIMES OF INDIA (Aug. 23, 2018), <http://timesofindia.indiatimes.com/articleshow/65508688.cms> [<https://perma.cc/6CRJ-G6LC>].

¹⁷⁶ Avelar, *supra* note 16; Mike Isaac & Kevin Roose, *Disinformation Spreads on WhatsApp Ahead of Brazilian Election*, N.Y. Times (Oct. 19, 2018), <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html> [<https://perma.cc/4QDF-SRPU>]; Cristina Tardáguila, Fabrício Benevenuto & Pablo Ortellado, *Opinion, Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It*, N.Y. Times (Oct. 17, 2018), <https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html> [<https://perma.cc/AB5Q-YG7W>].

¹⁷⁷ See Anthony Boadle, *Facebook’s WhatsApp Flooded with Fake News in Brazil Election*, REUTERS (Oct. 20, 2018, 12:33 PM), <https://www.reuters.com/article/us-brazil-election-whatsapp-explainer/facebooks-whatsapp-flooded-with-fake-news-in-brazil-election-idUSKCN1MU0UP> [<https://perma.cc/EKD4-AWF6>]; see also Avelar, *supra* note 16.

¹⁷⁸ See Bengani, *supra* note 15; Findlay & Schipani, *supra* note 15; Hitchen et al., *supra* note 15.

advocacy groups to apply the same limitation on forwarding that had been applied in India to WhatsApp users in Brazil, the company refused to do so, leaving the limit on simultaneous forwarding at twenty chats at a time.¹⁷⁹ In January 2019, WhatsApp declared its India experiment a success and applied the stricter limitations worldwide,¹⁸⁰ and reports found that this was effective in slowing the spread of false news.¹⁸¹ By this time, India was in the process of a general election, where false news spread on WhatsApp continued to be a grave concern.¹⁸² In August 2019 WhatsApp introduced a new feature: an icon displaying double arrows along with the label, “forwarded many times,” on all messages forwarded more than five times.¹⁸³

2020 brought another category of false news to WhatsApp – news about the COVID-19 epidemic. In fact, false news around the virus began almost immediately and spread alongside the virus itself.¹⁸⁴ As was the

¹⁷⁹ See Brad Haynes & Anthony Boadle, *Despite Brazil Election Turmoil, Facebook Stands by WhatsApp Limits*, Reuters (Oct. 23, 2018, 1:25 PM), <https://www.reuters.com/article/us-brazil-election-idINKCN1MX1W2> [<https://perma.cc/8D48-R4KU>]; see also Melo et al., *supra* note 13, at 376; Shashank Bengali, *How WhatsApp is Battling Misinformation in India, Where “Fake News is Part of Our Culture,”* L.A. Times (Feb. 4, 2019, 4:00 AM PST), <https://www.latimes.com/world/la-fg-india-whatsapp-2019-story.html> [<https://perma.cc/XE3J-ZW6P>].

¹⁸⁰ Rishi Iyengar, *WhatsApp Tightens Limit on the Number of People You Can Share Messages With*, CNN BUS. (Jan. 21, 2019), <https://www.cnn.com/2019/01/21/tech/whatsapp-forwarding-limits-india/index.html> [<https://perma.cc/944X-LW6E>]; see *More Changes to Forwarding*, *supra* note 173.

¹⁸¹ See Angela Chen, *Limiting Message Forwarding on WhatsApp Helped Slow Disinformation*, MIT TECH. REV. (Sept. 26, 2019), <https://www.technologyreview.com/2019/09/26/434/whatsapp-disinformation-message-forwarding-politics-technology-brazil-india-election/> [<https://perma.cc/75R8-WNFJ>].

¹⁸² See Billy Perrigo, *How Volunteers for India’s Ruling Party Are Using WhatsApp to Fuel Fake News Ahead of Elections*, TIME (Jan. 25, 2019, 7:48 AM EST), <https://time.com/5512032/whatsapp-india-election-2019/> [<https://perma.cc/C94H-VTJ5>]; Ponniah, *supra* note 17.

¹⁸³ *About Forwarding Limits*, *supra* note 19; Pranav Dixit, *WhatsApp Is Now Letting Users Know When a Message Has Been Forwarded Too Many Times*, BUZZFEED NEWS (Aug. 2, 2019, 5:34 AM), <https://www.buzzfeednews.com/article/pranavdixit/whatsapp-double-arrows-forwarded-messages-misinformation> [<https://perma.cc/LL7U-SA92>].

¹⁸⁴ Gerrit De Vynck, Riley Griffin & Alyza Sebenius, *Coronavirus Misinformation Is Spreading All Over Social Media*, BLOOMBERG (Jan. 29, 2020, 4:09 PM EST), <https://www.bloomberg.com/news/articles/2020-01-29/coronavirus-misinformation-is-incubating-all-over-social-media> [<https://perma.cc/9CCV-6VX2>].

case with other social media platforms, WhatsApp was broadly used to propagate false information about COVID, its sources, how to prevent or treat it, as well as harmful information about the vaccines.¹⁸⁵ WhatsApp quickly took notice of this worrisome trend and in April 2020 announced another change to its forwarding mechanism.¹⁸⁶ Building on its previously introduced category of messages “forwarded many times” WhatsApp added another restriction on such messages, cutting the number of times they could be forwarded from five to only a single chat.¹⁸⁷

In April 2020, WhatsApp reported that these restrictions (and others) had the effect of lowering the spread of viral messaging by seventy

¹⁸⁵ See K.J. Kevin Feng, Kevin Song, Kejing Li, Oishee Chakrabarti & Marshini Chetty, *Investigating How University Students in the United States Encounter and Deal With Misinformation in Private WhatsApp Chats During COVID-19*, Proceedings of the Eighteenth Symposium on Usable Privacy and Security 427-38 (2022); Antônio Diogo Forte Martins, Lucas Cabral, Pedro Jorge Chaves Mourão, José Maria Monteiro & Javam Machado, *Detection of Misinformation About COVID-19 in Brazilian Portuguese WhatsApp Messages*, in International Conference on Applications of Natural Language to Information Systems 199, 199 (2021); Masroor Ahmed, Muhammad Qamar Riaz, Munazza Qamar & Rohail Asghar, *Fake News Shared on WhatsApp During Covid-19: An Analysis of Groups and Statuses in Pakistan*, 17 Media Educ. 4, 4 (2021); Jeremy Bowles, Horacio Larreguy & Shelley Liu, *Countering Misinformation via WhatsApp: Preliminary Evidence from the COVID-19 Pandemic in Zimbabwe*, 15 PLOS One, Oct. 14, 2020, at 1; Cheesman et al., *supra* note 15, at 157; Carlos Elías & Daniel Catalan-Matamoros, *Coronavirus in Spain: Fear of “Official” Fake News Boosts WhatsApp and Alternative Sources*, 8 Media & Commc’n 462, 464 (2020); Carolina Moreno-Castro, Empar Vengut-Climent, Lorena Cano-Orón & Isabel Mendoza-Poudereux, *Exploratory Study of the Hoaxes Spread via WhatsApp in Spain to Prevent and/or Cure COVID-19*, 35 Gaceta Sanitaria 534, 535 (2021).

¹⁸⁶ *Keeping WhatsApp Personal and Private*, WHATSAPP BLOG (Apr. 17, 2020), <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private> [<https://perma.cc/6RRA-V2UV>] [hereinafter *Keeping WhatsApp Personal and Private*].

¹⁸⁷ Pranav Dixit, *WhatsApp Is Imposing Stricter Limits on Forwarding Messages to Slow Down Coronavirus Misinformation*, BUZZFEED NEWS (Apr. 7, 2020, 1:38 AM), <https://www.buzzfeednews.com/article/pranavdixit/coronavirus-whatsapp-forwarding-messages> [<https://perma.cc/4RGH-V2YN>]; Damola Durosomo, *Rejoice! WhatsApp Places New Limits on Chain Messages to Fight Fake News*, OKAYAFRICA (Apr. 7, 2020), <https://www.okayafrika.com/whatsapp-places-new-restrictions-on-message-forwarding-to-fight-fake-news-in-africa/> [<https://perma.cc/2V98-CMXZ>].

percent.¹⁸⁸ Since the content on WhatsApp is end-to-end encrypted there are several structural limitations in interpreting this piece of data. First, there is no way for us to ascertain whether the limitation in virality did indeed reach seventy percent. We therefore quote WhatsApp's findings in this regard. Furthermore, it is not possible to determine the nature of the content affected by the friction. We discuss this limitation further in Part V.C.

WhatsApp announced that it would further limit the rules for forwarding messages that had previously been forwarded but not yet frequently forwarded (i.e., those marked with a single forwarding arrow). Previously, it had limited users to forwarding such content to only five chats. Now, WhatsApp has added the restriction that such messages could be forwarded to only one group at a time.¹⁸⁹

3. The World Responds Positively

In the almost five years since WhatsApp began implementing limits on message forwarding, these moves have been well-received as a meaningful approach for limiting the spread of false news.¹⁹⁰ One

¹⁸⁸ Jon Porter, *WhatsApp Says Its Forwarding Limits Have Cut the Spread of Viral Messages by 70 Percent*, THE VERGE (Apr. 27, 2020, 5:28 AM PDT), <https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline> [<https://perma.cc/33FS-KLQ9>]; Marianna Spring, *Coronavirus: Viral WhatsApp Messages "Drop 70%"*, BBC NEWS (Apr. 27, 2020), <https://www.bbc.com/news/technology-52441202> [<https://perma.cc/6MQ3-E7D4>].

¹⁸⁹ Subin B., *WhatsApp to Introduce New Forwarding Limits to Reduce Spam*, BEEBOM (Apr. 2, 2022, 2:21 PM), <https://beebom.com/whatsapp-new-forwarding-limits-reduce-spam/> [<https://perma.cc/C4UX-T9J2>]; *WhatsApp Is Enabling New Forwarding Limits*, WABETAINFO (Apr. 1, 2022), <https://wabetainfo.com/whatsapp-is-enabling-new-forwarding-limits/> [<https://perma.cc/JJ7T-FZZU>]; Christian Zibreg, *WhatsApp Limits Message Forwards to Other Groups to Curb Misinformation*, IDOWNLOADBLOG (Apr. 4, 2022), <https://www.idownloadblog.com/2022/04/04/whatsapp-forward-message-group-chat-limit-tests/> [<https://perma.cc/N6Q9-93AZ>].

¹⁹⁰ See, e.g., Chinmayi Arun, *On WhatsApp, Rumours, and Lynchings*, 54 ECON. & POL. WKLY. 30, 35 (2019) ("On the face of it, this looks like it may make a difference. . ."); Hadas Gold, *WhatsApp Tightens Limits on Message Forwarding to Counter Coronavirus Misinformation*, CNN BUS. (Apr. 7, 2020, 2:07 PM EST), <http://edition.cnn.com/2020/04/07/tech/whatsapp-misinformation-forward-limit/index.html> [<https://perma.cc/M9RQ-6FKT>] ("Experts welcomed the tighter limit announced on Tuesday but said it still doesn't go far enough."); Kurt Wagner, *WhatsApp Will Drastically Limit Forwarding Across*

reporter even heralded these changes as “putting truth and fiction on a more even footing.”¹⁹¹ Many expressed hopes that these changes would play a significant role in limiting the harmful spread of false news.¹⁹² A similar set of forwarding restrictions has also been implemented by Meta in its Facebook Messenger App.¹⁹³ These accolades have been a rare bit of good news for WhatsApp and its parent company Meta, who have reeled from bad press during the same time period, especially in late 2021 with the whistleblowing reports by Frances Haugen.¹⁹⁴

Praise has come from academics as well. A team of Brazilian computer scientists analyzed the forwarding restrictions imposed by WhatsApp using an epidemiological model, concluding that WhatsApp’s restrictions can reduce a measure of the speed with which a message propagates, which they dub the “velocity of dissemination,” by one order of magnitude.¹⁹⁵ Researchers at the London School of Economics

the Globe to Stop the Spread of Fake News, Following Violence in India and Myanmar, VOX (July 19, 2018, 11:44 PM EST), <https://www.vox.com/2018/7/19/17594156/whatsapp-limit-forwarding-fake-news-violence-india-myanmar> [<https://perma.cc/K2TY-H6JK>] (“Limiting the rate at which people can forward messages won’t solve the problem, of course, but WhatsApp hopes it will slow down the viral impact that social networks have become known for.”).

¹⁹¹ Casey Newton, *WhatsApp Puts New Limits on the Forwarding of Viral Messages*, THE VERGE (Apr. 7, 2020: 12:00 AM PST), <https://www.theverge.com/2020/4/7/21211371/whatsapp-message-forwarding-limits-misinformation-coronavirus-india> [<https://perma.cc/53WP-AMAK>].

¹⁹² See, e.g., Melo et al., *supra* note 13, at 11 (finding that “[the] limits imposed on message forwarding and broadcasting (e.g. up to five forwards) offer a delay in the message propagation of up to two orders of magnitude . . .”).

¹⁹³ Jay Sullivan, *Introducing a Forwarding Limit on Messenger*, META NEWSROOM (Sept. 3, 2020), <https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/> [<https://perma.cc/GBR3-SBQ9>].

¹⁹⁴ E.g., Jeff Horwitz, *The Facebook Whistleblower, Frances Haugen, Says She Wants to Fix the Company, Not Harm It*, WALL ST. J. (Oct. 3, 2021, 7:36 PM EST), <https://www.wsj.com/articles/facebook-whistleblower-frances-haugen-says-she-wants-to-fix-the-company-not-harm-it-11633304122> [<https://perma.cc/57J2-TZH3>] (profiling Facebook whistleblower Frances Haugen); Georgia Wells, Jeff Horwitz & Deepa Seetharaman, *Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show*, WALL ST. J. (Sept. 14, 2021, 7:59 AM EST), <https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739> [<https://perma.cc/3DZZ-QU22>] (reporting that Facebook knew that Instagram caused mental health issues).

¹⁹⁵ Melo et al., *supra* note 13, at 381.

described the various steps WhatsApp took in India to counter false news including: an effort to identify and ban accounts sharing bulk messages, the design and implementation of digital literacy efforts, advertising tips for identifying false news in local newspapers, and operating a tipline for users to forward messages suspected of including false news.¹⁹⁶ While they regarded some of these efforts as ineffective, they described the restriction on forwarding as a positive (though imperfect) attempt at limiting the spread of false news, criticizing how easy and risk-free it had been to bulk-forward false news before these restrictions were implemented.¹⁹⁷ Notably, all of these studies took the efficacy of WhatsApp's changes for granted, and no academic study we could find tested whether any were improperly implemented or whether the limits could be circumvented. It is also important to note that the research analyzing the efficacy of the restrictions involved situations that were very different than the original situation which WhatsApp was initially using forwarding limitations to overcome. As described above, WhatsApp was called upon by the Indian government to limit the spread of conspiracy theories as perpetuated by villagers, many of whom were using their first cell phone to spread the false news.¹⁹⁸ While this type of friction may have been effective in limiting false news in these circumstances, it may not easily translate in situations involving users with more experience with this kind of technology. Despite this somewhat unique setting in which it was introduced, the small, emerging community of legal scholars of friction has held WhatsApp as a paradigmatic example of the kind of approach that should be copied, emulated, and expanded in other contexts as well. Frischmann and Benesch mention WhatsApp's restriction on forwarding as an example of their friction-in-design approach.¹⁹⁹ They describe this type of friction as "tak[ing] the form of a delay and extra steps and effort to reach a larger audience."²⁰⁰ Recognizing that WhatsApp's forwarding restriction is content-neutral, they even go so far as to recommend that

¹⁹⁶ BANAJI ET AL., *supra* note 138, at 20.

¹⁹⁷ *See id.* at 22-23.

¹⁹⁸ *How WhatsApp Helped Turn an Indian Village into a Lynch Mob*, *supra* note 168; see discussion *supra* Part II.B.

¹⁹⁹ Frischmann & Benesch, *supra* note 9, at 413.

²⁰⁰ *Id.*

governments could draw inspiration from WhatsApp to impose similar restrictions.²⁰¹ Ellen Goodman also cites the WhatsApp example favorably, highlighting that the friction introduced by the platform creates “higher cognitive and logistical burdens on those who would amplify the noise.”²⁰² The overall sentiment expressed by all has been that WhatsApp has found an effective way to address the ease with which users could share false news on the platform.

4. The Surprising Amount We Do Not Know

We undertook this project expecting to highlight the forwarding restrictions imposed by WhatsApp alongside two or three other similar examples of the tech industry’s use of friction to combat false news to consider how regulators might encourage other platforms to follow suit. We had planned to devote a small discussion, no more than a few paragraphs, to explaining exactly what WhatsApp had done and how effective it had been, based on the large, accumulated base of knowledge of what others had already learned or reported.

What we found surprised us and led our work in a new direction. The reporting on exactly what WhatsApp had done was surprisingly thin, void of detail, and overly credulous about the company’s own descriptions.²⁰³ Academic researchers had done a bit more to look under the hood, although most of these studies had been conducted by social

²⁰¹ *Id.* at 441.

²⁰² Goodman, *supra* note 4, at 649.

²⁰³ See Gowen & Dwoskin, *supra* note 165; Daniel Funke, *WhatsApp Is Limiting Message Forwarding to Cut Down on Fake News*, POYNTER (July 20, 2018), <https://www.poynter.org/fact-checking/2018/whatsapp-is-limiting-message-forwarding-to-cut-down-on-fake-news/> [<https://perma.cc/CZ5U-35YF>]; Andrew Hutchinson, *WhatsApp Implements New Restrictions on Message Forwarding to Limit the Spread of Misinformation*, SOC. MEDIA TODAY (Apr. 8, 2020), <https://www.socialmediatoday.com/news/whatsapp-implements-new-restrictions-on-message-forwarding-to-limit-the-spr/575654/> [<https://perma.cc/EGL2-MX2H>]; David Ingram, *WhatsApp Limits Forwarding of Messages to Try to Slow Coronavirus Misinformation*, NBC NEWS (Apr. 7, 2020, 11:31 AM PST), <https://www.nbcnews.com/tech/tech-news/whatsapp-limits-forwarding-messages-try-slow-coronavirus-misinformation-n1178606> [<https://perma.cc/ZS66-AEWN>]; Hamza Shaban, *WhatsApp Is Trying to Clamp Down on Viral Misinformation with a Messaging Limit*, WASH. POST (Jan. 22, 2019, 10:19 AM EST), <https://www.washingtonpost.com/technology/2019/01/22/whatsapp-is-trying-clamp-down-viral-misinformation-with-messaging-limit/> [<https://perma.cc/PLH2-YWBX>].

scientists who took for granted what WhatsApp had claimed they had done rather than by technical experts with the tools to investigate how the restriction itself has been implemented.²⁰⁴

In the end, we — a team of researchers trained in both computer science and law — have conducted what we believe is the most thorough academic study to date of WhatsApp’s limits on forwarded and frequently forwarded messages. Our results will serve at least three purposes. First, they will help policymakers and the public understand previously unknown caveats to this system, second guessing the conventional wisdom that WhatsApp has been effective at limiting forwarding and possibly at combatting the spread of false news. Second, they will help false news researchers better understand an oft-cited example, perhaps spurring new research questions. We pose some of those questions in Part V. Finally, they will help the emerging community of friction scholars by revealing the types of tradeoffs that we think are often encountered in friction-focused solutions.

We present our technical results in the next Part. Before we do, we want to clarify two important points.

First, at least based on what we know now, we are not sharply critical of the disconnect between what WhatsApp and Meta have claimed to have done and the reality of what they have done. We cannot point to any single obvious misstatement from the companies, and we are not calling on the FTC to open a section five deception investigation.²⁰⁵ WhatsApp, at worst, was silent about the ease with which their limits

²⁰⁴ See, e.g., Lucía-Pilar Cancelas-Ouviña, *Humor in Times of COVID-19 in Spain: Viewing Coronavirus Through Memes Disseminated via WhatsApp*, 12 *Frontiers Psych.* 1, 1 (2021) (analyzing memes spread through WhatsApp during COVID in Spain); Feng et al., *supra* note 185, at 427 (identifying WhatsApp as a central source of information during COVID); Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Davin Gaffney & Scott A. Hale, *Research Note: Tiplines to Uncover Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian General Election on WhatsApp*, 3 *Harv. Kennedy Sch. Misinfo. Rev.*, 2022, at 1 (analyzing the effectiveness of a crowd sourced tipline in limiting the spread of disinformation); Melo et al., *supra* note 13, at 372 (proposing a methodology to test the effectiveness of forwarding limitations on the spread of disinformation); Sananda Sahoo, *Political Posters Reveal a Tension in WhatsApp Platform Design: An Analysis of Digital Images from India’s 2019 Elections*, 23 *Television & News Media* 874, 874 (2022) (examining the effects of WhatsApp as a mode of dissemination of political posters around the time of the Indian elections).

²⁰⁵ See 15 U.S.C. § 45.

could be circumvented, and their silence may have led observers — including experts — to think the system is more robust than it is. In Part V, we will talk about how friction solutions sometimes encounter arms races and superusers, and how friction implementers need to make reasoned decisions about how far to engage in those fights.

Second, although our results reveal how WhatsApp’s forwarding limits work, and how easily they can be circumvented, we have not in this study measured how effective the tools have turned out to be in either combatting virality or false news. We have not tried to duplicate WhatsApp’s self-declared “70%” figure at reducing message virality,²⁰⁶ and due to the end-to-end-encrypted nature of the service, neither we nor anybody else could determine how effective the tool is at combatting false news, at least not without surmounting many significant technical challenges. We consider all of these in Part V and the Conclusion, which lay out a research agenda for future work.

IV. TECHNICAL INVESTIGATION OF WHATSAPP’S FRICTION

In this part we describe the two types of technical analysis of WhatsApp we conducted. First, we have verified the restrictions on the user level, i.e. what users see and what they can and cannot do. Second, we checked the underlying code of the app and discovered previously unknown limitations about the way these restrictions have been implemented.

A. Methodology

We followed a multi-step methodology, using tools and techniques that have been well-established in computer science and computer security research, and that have been adopted in some legal scholarship.²⁰⁷

²⁰⁶ Porter, *supra* note 188.

²⁰⁷ For a discussion of static and dynamic analysis, see Cyrille Artho & Armin Biere, *Combined Static and Dynamic Analysis*, 131 ELEC. NOTES THEORETICAL COMPUT. SCI. 3, 3 (2005); Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 642-52 (2017). For examples of studies that use Chrome DevTools, see Matthias Heinrich, Franz Lehmann, Franz Josef Grüneberger, Thomas Springer & Martin Gaedke, *Analyzing the Suitability of Web Applications for a Single-User to Multi-User*

First, we subjected WhatsApp's current architecture to a simple but rigorous *behavioral analysis*. To more precisely understand the extent of what WhatsApp had done, we used the WhatsApp app to send messages back-and-forth between users and groups to discover the specific steps that would trigger a forbidden action and to document the way the app's user interface would communicate the restriction to the user.²⁰⁸

Second, we conducted a *static analysis* of the web-based version of the WhatsApp app.²⁰⁹ This means we inspected the JavaScript source code that implements the entire WhatsApp app in a web browser. We did not have access to the source code for Android or iOS, so we limited this and the dynamic analysis described next to only the web platform. From over 200,000 lines of JavaScript code, we isolated the handful of lines of code that kept track of how many times a message had been forwarded as well as the separate lines of code that prevented a user from some forwarding behaviors.²¹⁰ This gave us much deeper insight into the tradeoffs WhatsApp had chosen, and it gave us clues for our next phase of research.

Transformation, in PROCEEDINGS OF THE 22ND INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 249, 250 (2013) (uses Chrome DevTools breakpoints to perform dynamic analysis); Shaown Sarker, Jordan Jueckstock & Alexandros Kapravelos, *Hiding in Plain Site: Detecting JavaScript Obfuscation Through Concealed Browser API Usage*, in PROCEEDINGS OF THE ACM INTERNET CONFERENCE 648, 649 (2020) (uses DevTools for static analysis of website source code); Antoine Saverimoutou, Bertrand Mathieu & Sandrine Vaton, *Web Browsing Measurements: An Above-the-Fold Browser-Based Technique*, 2018 INT'L CONF. DISTRIBUTED COMPUTING SYS. 1630, 1632-34 (2018) (uses DevTools to analyze network logs and retrieve screenshots). For surveys that involve code inspection and/or static analysis of code repositories on GitHub, see Mohammad Gharehyazie, Baishakhi Ray & Vladimir Filkov, *Some from Here, Some from There: Cross-Project Code Reuse in GitHub*, in PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON MINING SOFTWARE REPOSITORIES 291, 291 (2017); Pamela H. Russell, Rachel L. Johnson, Shreyas Ananthan, Benjamin Harnke & Nichole E. Carlson, *A Large-Scale Analysis of Bioinformatics Code on GitHub*, 13 PLOS ONE, Oct. 31, 2018, at 1.

²⁰⁸ To capture variations in implementations, we used WhatsApp's official apps on Android, Apple iOS, and web-based platforms.

²⁰⁹ See Kroll et al., *supra* note 207, at 647-50.

²¹⁰ Counting lines of code is an imprecise undertaking. Like most web-based apps, the WhatsApp source code is *minified*, meaning it is presented compactly to save memory usage and download time. It was only after we "pretty printed" the code, meaning we unpacked it to make it easier for a human to understand, that we ended up with more than 200,000 lines of code.

Third, we conducted a *dynamic analysis*, studying the behavior of the web version of WhatsApp while running, including by altering the source code to see if we could find ways to circumvent the forwarding restrictions.²¹¹ This step confirmed what we had hypothesized during static analysis; we were able to prove that some of these controls are circumventable by making very minor changes to the code.

Finally, we used what we had learned to search the web for examples of circumvention in the wild. Our static and dynamic analyses had given us keywords that, if found in source code, would strongly suggest attempts to circumvent. We were able to find some compelling examples.

B. Behavioral Analysis

First, we attempted to verify the behavior claimed by WhatsApp on their website. WhatsApp claims:²¹²

You can forward a message with up to five chats at one time. If a message has already been forwarded, you can forward it to up to five chats, including a maximum of one group.

Messages forwarded through a chain of five or more chats, meaning it's at least five forwards away from its original sender, have a [double arrow] icon and "Forwarded many times" label displayed. These messages . . . can only be forwarded to one chat at a time. This helps keep conversations on WhatsApp intimate and personal. This also helps slow down the spread of rumors, viral messages, and fake news.²¹³

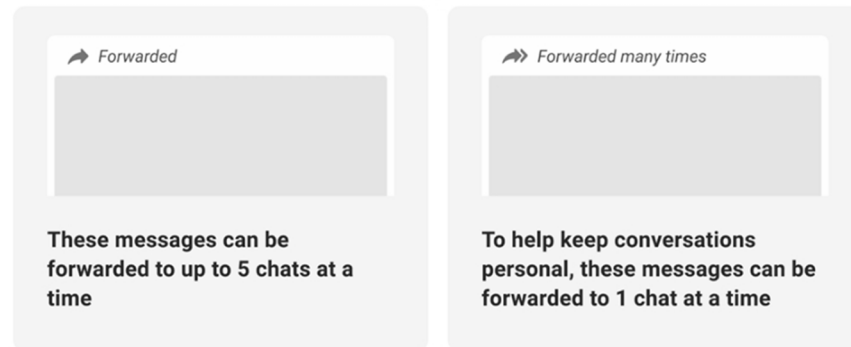
Elsewhere in the documentation,²¹⁴ the user is shown how to recognize both forwarded and frequently forwarded messages visually:

²¹¹ See Kroll et al., *supra* note 207, at 650-52.

²¹² *About Forwarding Limits*, *supra* note 19.

²¹³ *Id.*

²¹⁴ *Id.*



Amidst the dishonorable modern landscape of confusing user documentation,²¹⁵ these explanations are admirably concise and clear and, as we discovered through behavior testing, mostly accurate. We recognized quickly, however, that even before we started probing the engineering details, these descriptions left many questions unanswered: what does it mean for a message to have been forwarded? If one user forwards a message to five different people is that message considered “frequently forwarded,” or does that require five different people to forward the message *seriatim*? What limits apply to non-forwarded messages, meaning messages that have been typed in directly? Can a user circumvent any of these limits by copying-and-pasting a message rather than forwarding?

1. Behavioral Analysis Explained

We tested the behavior using brand-new accounts created for this purpose and three different devices to give us access to the iOS, Android, and desktop browser versions of WhatsApp.

From these devices, we sent simple text-only messages between the three devices in varying permutations. By identifying the situations in which we were prevented from taking an action and by documenting

²¹⁵ See, e.g., Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty & Arvind Narayanan, *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites*, 3 *PROC. ACM HUM.-COMPUT. INTERACTION*, Nov. 2019, at 1-2 (analyzing dark patterns, defining them as “user interface design choices that benefit an online service by coercing, steering, or deceiving users into making unintended and potentially harmful decisions” and using shopping websites as an example).

what the user interface presented to the user in those cases, we were able to write a detailed behavioral summary of how WhatsApp’s forwarding limits work today.

2. Behavioral Analysis Results: The User’s Point of View

We present these behavioral analysis results in great detail, going a bit further than needed to support our recommendations in Part V. We do so to help guide the source code analyses that follow, but also to fill a gap in the literature, giving other researchers an authoritative guide for how these restrictions operated at this moment in time. Consider this the “missing manual” describing the technical minutiae of these important rules.

When user Alice sends a message, she must first select her intended recipient or recipients. In WhatsApp parlance, she can send to only a single *chat*, meaning a single user or group. To that one chat, she can send text, images, files, or video. None of the WhatsApp friction-focused forwarding rules apply to the original sender of a message, to Alice. We say that Alice has created an *original message*, which is not the target of any of WhatsApp’s forwarding rules.

Alice’s original message may or may not be false news. Like all the rules we are describing in this Subsection, this entire system is content-neutral,²¹⁶ limiting the sharing of messages based only on non-content considerations involving what has happened to the message in the past. If Alice wants to author election misinformation or send a deepfake video, she is permitted to do so, and she can send it to one user, call him Bob, to multiple users subscribed to a single group, or by creating a new group with up to 1,024 users.²¹⁷

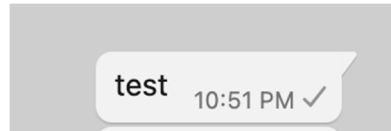
Alice can spread her false news far and wide, by sending the same original message repeatedly to multiple different groups in turn. Importantly, when Alice presses the “forward” button for this message, it retains its status as an original message and does not trigger the restrictions described below. Each new re-send requires only a few screen presses, allowing her to choose another group with up to 1,024

²¹⁶ See *infra* Part V.B (discussing content-neutrality).

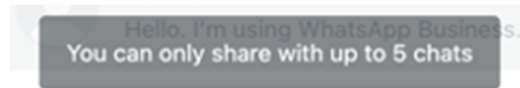
²¹⁷ *How to Create and Invite into a Group*, *supra* note 151.

people in a matter of seconds. She is limited only by her time and fortitude, and by the number of group links she can find.

Next consider Bob, one of the direct recipients of Alice's original message. In Bob's chat with Alice, Alice's original message appears in a little speech bubble, devoid of any additional icons. This is how Bob knows the message is an original message created by Alice. Here is a screenshot of a message in this state during our testing:



Bob is then permitted to *forward* Alice's message to other users or groups, and forwarding is what triggers all the rules we investigated. As a forwarder, Bob is subject to constraints that did not affect Alice. When Bob forwards a message, he can choose to send it to more than one chat, but as soon as he chooses a sixth chat as a destination, a pop-up message appears informing him that he can share only with up to five chats. The sixth chat is not permitted.



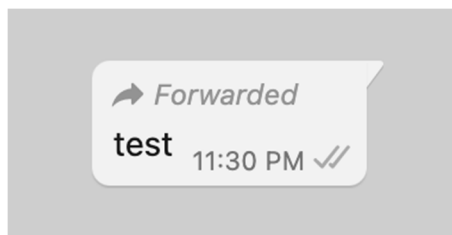
This is the first WhatsApp frictional constraint we have considered. It slows the forwarding of all messages by allowing Bob to forward a message to no more than five users or groups. Because groups can have as many as 1,024 users, this is effectively a 5,120-user cap on any single forward.

As with many friction-based solutions, Bob is slowed but not stopped from forwarding messages to even more people, as he can reforward Alice's original message as many times as he wants. With each reforward, he can select another five users or groups, meaning he is limited only by his time and fortitude, and by the number of group links he can find.

Turn next to Charlie, to whom Bob forwarded Alice's original message. Bob may have selected Charlie specifically, or Charlie may be a member of one of the groups selected by Bob. Charlie has received *forwarded* content because he did not receive the message directly from

Alice. Charlie and the message from Bob in Charlie's possession are focal points of WhatsApp's attempts to slow down the spread of false news, and Charlie experiences additional friction with respect to this message in several ways.

First, when Charlie views the forwarded message, he sees it annotated with an icon, a small arrow indicating that this is not an original message. Next to the icon is the label, "Forwarded." Because the message is only one hop away from its original source, from Alice, only one arrow is shown.²¹⁸ Again, a screenshot from our tests:



If Charlie chooses to forward Bob's message, a forward of Alice's original message, Charlie will encounter greater limits than those that faced Bob. Like Bob, Charlie will be allowed to select only five chats as destinations. But among those five, Charlie will be permitted to select only a single group. As soon as Charlie selects a sixth destination, he will encounter a message that says, "You can only share with up to 5 chats." As soon as Charlie selects a second group, whether or not he has selected five chats, he will encounter a message that says, "Forwarded messages can only be sent to one group chat at a time." Whereas Bob was permitted to send to 5,120 people at a time, Charlie is limited to no more than 1,028, four individuals and one group of up to 1,024.

Like Bob, Charlie can reforward Bob's forward of Alice's message as many times as Charlie wants. Each time, he is permitted to select a new list of up to 1,028 people.

Next, consider Carlos, another of Bob's forwarded message recipients. Carlos sees exactly what Charlie sees and is subject to exactly the same constraints. Importantly, nothing Charlie does with Bob's

²¹⁸ This marking is similar to the "Fwd:" text appearing in forwarded emails. Unlike email text, the "forwarded arrow" cannot simply be deleted before forwarding.

messages affects what Carlos sees or can do. This goes also for Carol, Carmine, Cleo, and every other one of Bob's recipients.

This is an important takeaway about the choices and tradeoffs WhatsApp made when implementing friction. WhatsApp applies restrictions on forwarding depending on how many hops away from the author of the original message a user is. Another option would have been to restrict forwarding depending on how many recipients received the original message. For example, a message forwarded by Bob to Charlie and Carlos could have been considered to have been forwarded more times than the message Bob sent exclusively to Charlie. A computer scientist might say that WhatsApp measures the depth but not breadth with which a message has been shared in its friction-focused calculations.

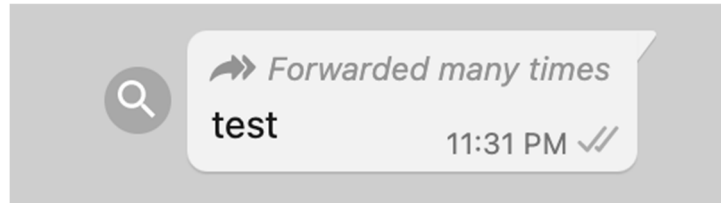
Return to Bob for a moment. Bob triggered WhatsApp's forwarding rules by forwarding Alice's original message. What if Bob had wanted to circumvent those limits? For example, Bob could have copied the text, image, file, or video — the message's content — from Alice's message to his device's clipboard or file storage. He could have then re-pasted or attached that content into a new message. Our tests confirm that at this point, the rules imposed on forwarding messages no longer apply.²¹⁹ WhatsApp will treat this message as an original message. Bob can send it to as many groups or users as he likes, albeit one group or user at a time, just like Alice. More importantly, Charlie (and Carlos, Cleo, etc.) will now receive the message as an original rather than forwarded message. The message will not be annotated with an arrow. And Charlie can resend it to up to 5,120 rather than only 1,028 users at a time.

Continue to follow a message forwarded down successive hops away from Alice. If Bob forwards to Charlie who forwards to Delilah who forwards to Eleanor who forwards to Felix, everyone in the chain of forwards starting with Charlie will see exactly what Charlie saw and be subjected to the same rules imposed on Charlie. The message will

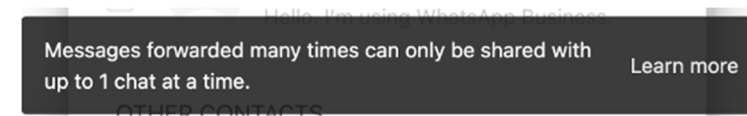
²¹⁹ Sandra Gutierrez G., *How to Get Rid of the "Forwarded" Label on WhatsApp*, POPULAR SCI. (Aug. 10, 2021, 2:14 PM EST), <https://www.popsoci.com/diy/forward-messages-whatsapp/> [<https://perma.cc/7X2T-3PL4>] (providing a how-to guide on using cut-and-paste to circumvent forwarding limits without reflecting at all on any potential implications of doing so).

appear with a single arrow, and its recipient will be allowed to reforward to no more than five chats, only one of which may be a group.

It's when Felix, who is six hops away from Alice and five hops away from Bob, forwards to Gita that we encounter WhatsApp's final set of restrictions. First, Gita will see a new icon, a set of doubled arrows. Next to the icon will be the label, "Forwarded many times."²²⁰ From our tests:



When Gita tries to forward the message onward, she'll be subject to stricter limits than those imposed on Bob, Charlie, Delilah, Eleanor, and Felix. Gita is permitted to forward the message to only one chat at a time, either a single group or a single user. Pressing a second chat will result in the message, "Messages forwarded many times can only be shared with up to 1 chat at a time." Since a single group can have up to 1,024 users, this reduces the numeric limit from 1,028 to 1,024.






After Gita, the same frequently forwarded treatment will apply to every other recipient. Frequently forwarded messages can be sent to only a single chat — group or user — at a time.

The following table summarizes our behavioral observations:

User	Hops from Alice	Forwards away from Bob	Icon	Forwarding Limits	
				Total number of chats	Total number of groups
Alice	n/a	n/a	None	n/a	n/a

²²⁰ On some versions of WhatsApp, a "search" icon will appear next to the message, and clicking on it will elicit a pop-up asking the user to consider the veracity of the message. *Search the Web*, WHATSAPP BLOG (Aug. 3, 2020), <https://blog.whatsapp.com/search-the-web> [<https://perma.cc/ZL4N-Y87J>].

Bob	1	n/a	None	5	5
Charlie (Carlos, Cleo, etc.)	2	1		5	1
Delilah	3	2		5	1
Gita	6	5		1	1

C. Static Analysis of Source Code

1. Static Analysis

Armed with this detailed understanding of how users experience WhatsApp's forwarding limits, we next explored the engineering decisions made in implementing these rules. We suspected that doing so would help us understand some of the tradeoffs WhatsApp made in implementing friction, which we hoped would inform regulators considering mandating similar restrictions.

First, we conducted *static analysis*, a well-known technique in computer security research that entails analyzing the code underlying an app.²²¹ A researcher engaged in static analysis traces the lines of code to try to reveal an app's underlying logic and operation. Static analysis is contrasted with dynamic analysis, in which the researcher launches the app and inspects how it operates in action.²²²

We focused both our static and dynamic analyses on the web-based version of the official WhatsApp app because this made it much easier to access the source code.²²³ Like most modern interactive web-based

²²¹ See Kroll et al., *supra* note 207, at 647-50.

²²² See *id.* at 650-52. As a rough analogy, static analysis is like analyzing the various parts of a catapult, such as the length of its arm and the tensile strength of the metal used in its spring, to try to predict how far it will launch a projectile, while dynamic analysis would be like observing the movement of the arm and measuring the projectile in flight.

²²³ There are also two types of static analysis, depending on whether the researcher can access the app's source code or instead be forced to content themselves with object code. The difference is a matter of degree of comprehensibility and thus a difference in

apps, the web version of WhatsApp is written in JavaScript, a language built into all modern browsers.²²⁴ JavaScript is delivered as source code, in all of its expressive readability; only once it is loaded into a web browser does it get interpreted into a machine executable format.²²⁵ Even better, all modern browsers provide rich tools for revealing and manipulating the JavaScript delivered on the web.²²⁶ We used the tools built into the Google Chrome browser (known as “Chrome DevTools”) for most of the static and dynamic analyses described below.

This points to a caveat of this work: we tested the web version of WhatsApp exclusively, and we cannot yet confirm that any of our findings hold true for the Apple or Android versions of the WhatsApp app.

difficulty. Source code is written by humans, using relatively expressive language that is understandable to humans. Source code is almost always rendered into equivalent object code, the raw 1’s-and-0’s binary data that a computer can run. Object code is stripped of most human-friendly language, in order to optimize how quickly it runs. The result is that object code requires much more specialty knowledge and painstaking reverse engineering to comprehend than source code. Apps designed to run in Apple and Android smartphones are usually compiled, meaning rendered into object code, before installed on a device. The apps one downloads from the Google Play store or Apple App store do not include the source code, which is often a closely-held trade secret of the developer. We thus did not engage in static analysis of the apps found on smartphones. We could have decompiled the object code into equivalent source code. See Kevin Burk, Fabio Pagani, Christopher Kruegel & Giovanni Vigna, *Decompersion: How Humans Decompile and What We Can Learn From It*, in *PROCEEDINGS OF THE 31ST USENIX SECURITY SYMPOSIUM* 2765, 2765 (2022). We leave this checking step for future work.

²²⁴ See *JavaScript — Dynamic Client-Side Scripting*, MDN WEB DOCS, <https://developer.mozilla.org/en-US/docs/Learn/JavaScript> (last visited July 13, 2023) [<https://perma.cc/DAP7-PYFR>]. We verified that WhatsApp Web uses JavaScript by looking at the source code on the browser.

²²⁵ To be precise, JavaScript in a web browser is turned “just in time” into executable object code. See Jan Kasper Martinsen, Håkan Grahn & Anders Isberg, *A Comparative Evaluation of JavaScript Execution Behavior*, in *11TH INTERNATIONAL CONFERENCE ON WEB ENGINEERING* 399, 399 (Soren Auer, Oscar Diaz & George A. Papadopoulos eds., 2011).

²²⁶ See *Debugging in the Browser*, JAVASCRIPT.INFO (June 26, 2022), <https://javascript.info/debugging-chrome> [<https://perma.cc/TC9F-X9NY>]; *What Are Browser Developer Tools?*, MDN WEB DOCS https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Tools_and_setup/What_are_browser_developer_tools (last updated July 7, 2023) [<https://perma.cc/2T4N-BS7V>].

For our static analysis, we visited the URL <https://web.whatsapp.com/> using Google Chrome and logged into a WhatsApp account.²²⁷ Using Chrome DevTools, we were able to view, search through, and even manipulate the hundreds of thousands of lines of source code underlying WhatsApp.²²⁸ We then performed a keyword search through the source code for the word “forward,” then manually reviewed the results in order to find code that appeared to be involved in tracking forwarding.

2. Static Analysis Results

Our static analysis reveals that, under the hood, WhatsApp appears to track each message with a critical piece of metadata, assigned the name *forwardingScore*. Based on the name and the way in which this metadata is used, we guessed that it was used to track the number of forwarded hops a message is away from an original message, a guess we were able to later confirm during dynamic analysis.

For example, in a file called *app.js*, we found a function named *_forwardMessageAndSendToChat*. The name suggests that this is a critical function used when a message is forwarded, and in the function, it appears to gather metadata about a message, including this critical line:

```
197510 | forwardedFromWeb: !0,
197511 | forwardingScore: e.getForwardingScoreWhenForwarded(),
197512 | multicast: n
```

Line 197,511 in turn calls another function called *getForwardingScoreWhenForwarded*.²²⁹ This function reads as follows:

²²⁷ Given the end-to-end encrypted nature of WhatsApp and that it identifies accounts by phone, one must first log in to WhatsApp from a smartphone and then connect the web-based WhatsApp to that phone and account by scanning a QR code.

²²⁸ Most modern browsers come with similar capabilities. We could have used another browser for this analysis. *E.g.*, *Open the Inspector*, FIREFOX SOURCE DOCS, https://firefox-source-docs.mozilla.org/devtools-user/page_inspector/how_to/open_the_inspector/index.html (last visited July 13, 2023) [<https://perma.cc/W9N9-YC8S>]; *Use the Developer Tools in the Develop Menu in Safari on Mac*, SAFARI USER GUIDE, <https://support.apple.com/guide/safari/use-the-developer-tools-in-the-develop-menu-sfr120948/mac> (last visited July 13, 2023) [<https://perma.cc/SBV2-46FB>]. See *supra* note 210 for a discussion on some subtleties involved in discussing the number of lines of code we analyzed.

²²⁹ See *supra* note 210 (explaining subtleties about the way we refer to source code line numbers).

```

205749     getForwardingScoreWhenForwarded() {
205750         const e = this.numTimesForwarded + (this.id.fromMe ? 0 : 1);
205751         return e >= J.ServerProps.frequentlyForwardedThreshold ? P.default.FREQUENTLY_FORWARDED_SENTINEL : e
205752     }

```

Finally, line 205,750 refers to a different value called *numTimesForwarded*, which we found referenced in these lines of code:

```

205052         this.numTimesForwarded = (0,
205053         y.derived)((function() {
205054             return this.forwardingScore ? this.forwardingScore || 0 : this.isForwarded ? 1 : 0
205055         }
205056         ), ["isForwarded", "forwardingScore"]),

```

Putting these together, the third snippet assigns *numTimesForwarded* a value based on a message’s *forwardingScore*, in line 205,054. Line 205,750 adds 1 to this value in the function *getForwardingScoreWhenForwarded*, which in turn is used to alter *numTimesForwarded*. In plain language, it appears that when a user forwards a message, this code increases the metadata field called *numTimesForwarded* by one.²³⁰

D. Dynamic Analysis of Running Source Code

1. Dynamic Analysis

There are limits to static analysis. While studying the source code alone, we cannot confirm that a function’s name is an accurate description for what the function does nor that a variable’s name is an apt description for what the variable stores.²³¹ Only by observing the code while running — by engaging in *dynamic analysis* — can we confirm the suspicions we developed during the static analysis phase.²³²

Once again, we used Chrome DevTools to inspect the JavaScript in the web version of WhatsApp to observe the code in action. We set *breakpoints* on particular lines in the source code. A breakpoint tells Chrome to halt the operation of the app at an indicated point — to stop all processing whenever the line is reached.

Once the code is stopped by a breakpoint, we can take three very powerful actions: (1) we can *step* through the code, line-by-line; (2) we

²³⁰ Note also that the mention of “this.id.fromMe” means that when a user forwards a message they originally wrote to a second recipient, it does not increase the “numTimesForwarded” value.

²³¹ These are human-selected labels, and a human software developer might choose a misleading name inadvertently or to mislead readers.

²³² See Kroll et al., *supra* note 207, at 650-52.

can query all of the data stored by the app at that moment of time, revealing the full state of the app; and (3) we can alter this data, allowing us to modify the running code. Ability Three is especially powerful because it allows us to run experiments on WhatsApp, allowing us to test whether any of WhatsApp’s forwarding restrictions were circumventable.

2. Dynamic Analysis Results

First, we tested the code snippets we found during static analysis to confirm that they did operate as we guessed. Here again is the code we thought was used to increase the metadata counter *forwardingScore* when a message was forwarded. We set a breakpoint as indicated by the blue flag on line 205,750:

```

205749 |
205750 |         getForwardingScoreWhenForwarded() {
205751 |             const e = this._numTimesForwarded + (this.id.fromMe ? 0 : 1);
205752 |             return e >= J.ServerProps.frequentlyForwardedThreshold ? P.default.FREQUENTLY_FORWARDED_SENTINEL : e

```

This breakpoint stopped the app as soon as a message was about to be forwarded: once a user selected a recipient and pressed the “forward” button, Chrome DevTools froze at this line of source code. This confirmed that the logic we found during static analysis was used to count how many times a message had been forwarded.

We next discovered that the system used to count the number of forwards was easily circumventable. To do so, we forwarded a frequently forwarded message (one like the message received by Gita above, five forwards away from Bob). As before, the app stopped at the breakpoint. From Chrome DevTools’ “Console” tab,²³³ we entered this line of JavaScript code:

```
e = 0;
```

We were testing whether *getForwardingScoreWhenForwarded* could be coaxed to forget that a message had been frequently forwarded. Sure enough, from the recipient’s computer, the message no longer appeared with the double arrows indicating the frequently forwarded state. Moreover, the recipient was able to forward the message subject only to

²³³ A breakpoint temporarily freezes but does not quit the app. The Console app allows us to manipulate the contents of memory at that frozen moment in time. Then, we can resume the app but now with the altered values in memory.

the limits on forwarded messages (five destination chats including one group) rather than the limits on frequently forwarded messages (only one chat). With a little technical know-how and some time searching through source code, we were able to bypass WhatsApp's attempt to label a message as frequently forwarded. We were not overly alarmed by this example because this seemed no easier to do than simply cutting-and-pasting the content of a message, which also had the effect of wiping out a message's frequently forwarded status.

Next, we added a breakpoint to this line of app.js:

```

71368     e.assertAttr("from", "s.whatsapp.net");
71369     const t = e.child("props")
71370       , r = {
71371         serverPropsVersion: t.attrInt("version")
71372       };
71373     return t.forEachChildWithTag("prop", (e=>{
71374       switch (e.attrString("name")) {
71375         case "web_ctwa_context_compose_enabled":
71376           r.ctwaContextCompose = 1 === e.attrInt("value");
71377           break;
71378         case "ctwa_context_enabled":
71379           r.ctwaContextRender = 1 === e.attrInt("value");
71380           break;

```

And we entered in the Console this line of code:

```
r.frequentlyForwardedMax = 10;
```

Recall that frequently forwarded messages are supposed to be able to be forwarded to only one chat. This line of code allowed us to circumvent this limit, permitting us to forward to more than one chat. Much more surprisingly, we were even able to circumvent the upper limit of five total chats, to which all forwarded messages are subject! In the example above, we arbitrarily chose the number ten, which let us send a frequently forwarded message to ten chats. In fact, when we pressed on an eleventh chat, we encountered this error message:

Messages forwarded many times can only be shared with up to 10 chats at a time. [Learn more](#)

Our change had even altered the error message, which now reported a constraint entirely of our invention. Note too that we could have set this value even higher, increasing the supposed limit of one message to 100 or 1,000 or even higher.

Both of the bypass examples above demonstrate our ability to circumvent the numerical limits placed on forwarding individual

messages. The second is especially powerful in the context of false news because it allows a user to forward a frequently forwarded message to every user in their address book.

We next explored if we could circumvent these limits in a different way, rather than by circumventing the numerical limits themselves, instead by stripping the frequently forwarded status for every message in every chat in our account. This too proved to be surprisingly easy. We added a breakpoint to this code:

```
190677 |           }, ["isForwarded", "forwardingScore"]),
190678 |           this.isFrequentlyForwarded = 0,
190679 |           g.derived)((function() {
190680 |             return this.numTimesForwarded >= C.default.FREQUENTLY_FORWARDED_SENTINEL
190681 |           })
190682 |           ), ["numTimesForwarded"]),
190683 |           this.eventType = 0,
190684 |           g.derived)((function() {
```

With the code frozen in this spot, we entered this in the Console:

```
this.isFrequentlyForwarded = (0, g.derived)((function() {return false}),
["numTimesForwarded"]);
```

This caused every single “frequently forwarded” message in every single chat in our account to change from the double-arrowed “Forwarded many times” to the single-arrowed “Forwarded.” Forwarding those altered messages confirmed that they were subject to the more liberal restrictions placed on forwarded messages (five chats and one group) rather than the stricter restrictions ones placed on frequently forwarded messages (only one chat).

Our final technique was perhaps the most complete and systematic. In the code that limited the number of recipients who could receive a forwarded message, we found a variable n , that appeared to be the critical value that is checked to detect when a user has tried to add more than the maximum number of chats or groups. We determined that we could wipe n at a particular, critical juncture in the code, depicted here:²³⁴

²³⁴ To do so, we took advantage of a “logpoint.” Unlike a breakpoint, a logpoint does not halt execution of code. Instead, it evaluates a given line of code every time a specific point in the program is reached.

```

131844         ~ ~ ~ \
131845         \       \
131846         ~ ~ ~ \
131847         \       \
131848         \       \
131849         \       \
131850         \       \
131851         \       \
131852         \       \
131853         \       \
131854         \       \
131855         \       \
131856         \       \
131857         \       \
131858         \       \
131859         \       \

```

```

    actionText: j.fbt_("Learn more", null, {
      hk: "176vVf"
    }),
    onAction: ()=>{0,
      f.openExternalLink)((0,
        p.getFrequentlyForwardedFaqUrl())
    });

```

```

Line 131851: Logpoint
n = null

```

```

    if (null != n)
      return void y.ToastManager.open((0,
        0.jsx)(_.Toast, {
          msg: n,
          action: a
        }), y.ToastPosition.CENTER)
    }
    i.setVal(e, t, a),

```

With this simple change, we could ignore all of WhatsApp’s forwarding limits. We could send any WhatsApp message — whether non-forwarded, forwarded, or frequently forwarded — to an unlimited number of chats and groups at one time. This change would persist for as long as the browser tab was kept open.

E. Why Circumvention Matters

1. Does Circumvention Matter?

Based on our findings, users with a fair amount of technical skill can circumvent many of WhatsApp’s restrictions on message forwarding. Although the ability to set breakpoints and enter code into a console window probably surpasses the technical abilities of most users, it is the kind of task that can be taught quickly to only moderately skilled users.

More importantly, any circumvention that can be done with breakpoints and lines of JavaScript can also be built into a user-friendly plug-in tool or an automated bot. Such a plug-in or bot would be able to render all frequently forwarded messages as merely forwarded and erase the limit on the number of chats to which a forwarded message could be reforwarded.

2. Is Anyone Circumventing These Restrictions?

Although our research revealed no other academic research discussing these techniques or features of the WhatsApp source code,

we hypothesized that other WhatsApp users had already discovered and were already using similar techniques.

To test this hypothesis, we searched public, online code repositories for the variable and function names we had identified. These searches reveal that developers are indeed modifying fields such as *forwardingScore* in the wild. A search for “forwardingScore” in the popular software code-hosting platform GitHub²³⁵ returns over 1,000 results.²³⁶ A search for the even more revealing “forwardingScore: 0”²³⁷ gives over 100 results, mostly what appear to be WhatsApp bots that automatically forward messages.²³⁸ In particular, we found one code repository on GitHub, “Bosco,” that is a template for creating a WhatsApp bot that sets *forwardingScore* to zero. This bot template has been “forked” over 1,200 times, meaning developers have at least started creating personalized bots based on this template over 1,200 times, all pointing to the presence of a sizable ecosystem of bots that bypass forwarding limits.²³⁹

Beyond this evidence of intentional circumvention, there is another source of potential circumvention: neglect. When most users think of WhatsApp, they picture the official apps developed by Meta that are

²³⁵ GitHub is a website that allows users and communities to share and collaborate on coding projects. See *GitHub's Products*, GITHUB DOCS, <https://docs.github.com/en/get-started/learning-about-github/githubs-products> (last visited July 13, 2023) [<https://perma.cc/VUV7-W63V>].

²³⁶ *Code Search Results*, GITHUB, <https://github.com/search?q=forwardingScore&type=code> (last visited July 23, 2023) [<https://perma.cc/VE7Q-H2DU>] (requires a GitHub account to access search results).

²³⁷ Code that merely mentions *forwardingScore* isn't necessarily circumventing any restrictions. Code that assigns *forwardingScore* to zero, which is what this search query reveals, is likely changing the count of how many times a message had been forwarded, meaning it is likely circumventing one of WhatsApp's restrictions.

²³⁸ Code search for *forwardingScore*, GITHUB, <https://github.com/search?q=%22forwardingScore%3A+0%22&type=code> (last visited July 23, 2023) [<https://perma.cc/VE7Q-H2DU>]. There are dozens of more results if searching for *forwardingScore* set to other arbitrary values, such as “forwardingScore: 1” or “forwardingScore: 2”. These would have the same effect of bypassing forwarding limits.

²³⁹ Pepesir, *Bosco Bot*, GITHUB <https://github.com/pepesir/Bosco> (last visited July 23, 2023) [<https://perma.cc/7YTD-B97S>]. To be clear, a fork represents the start of a new software development project. It is likely that many of the 1,200 forks did not end in a fully functional new WhatsApp forwarding app.

available for Apple iOS, Android, and web browsers. In reality, WhatsApp is an ecosystem of compatible apps beyond the official three that are all able to send and receive messages to other users, whether or not they are using an official WhatsApp app. Many developers who are not affiliated with Meta have created compatible third-party apps that can communicate with users using WhatsApp.²⁴⁰ WhatsApp notes on its website that its terms and conditions officially forbid the use of “unofficial apps” in its Terms of Service, but given the widespread availability of third-party apps, WhatsApp either cannot or has chosen not to enforce this restriction aggressively.²⁴¹

We have discovered that some of these third-party apps do not set a *forwardingScore* at all on messages their users send or forward. Similarly, many apps do not increment the *forwardingScore* with each re-forward.²⁴²

Even if WhatsApp is tolerating some third-party apps, why is it not forcing all apps to implement the forwarding limits? It may be that WhatsApp wants to preserve *backwards compatibility*, meaning it did not want older versions of the software to stop working. Additionally,

²⁴⁰ Elliot Nesbo, *Why Unofficial WhatsApp Apps Are a Security Risk*, MAKEUSEOF (Oct. 29, 2022), <https://www.makeuseof.com/unofficial-whatsapp-security-risk/> [<https://perma.cc/HF6H-QXR2>]; Efe Udin, *Unofficial WhatsApp Apps Are Gaining Massive Popularity in Some Regions*, GIZCHINA (Mar. 9, 2020), <https://www.gizchina.com/2020/03/09/unofficial-whatsapp-apps-are-gaining-massive-popularity-in-some-regions/> [<https://perma.cc/LPZ2-U22P>].

²⁴¹ *About Unofficial Apps*, WHATSAPP HELP CTR., <https://faq.whatsapp.com/1217634902127718/> (last visited July 13, 2023) [<https://perma.cc/6P32-ZA5K>] (“Unofficial apps are fake WhatsApp apps, developed by third-parties which violate our Terms of Service. If you use these apps[,] your privacy and security are at risk . . . [and] [y]our account might also be temporarily or permanently banned.”); *WhatsApp Terms of Service*, WHATSAPP (Jan. 4, 2021), <https://www.whatsapp.com/legal/terms-of-service> [<https://perma.cc/5FE8-L5TR>] (“You must not (or assist others to) directly, indirectly, through automated or other means . . . (a) reverse engineer, alter, modify, create derivative works from, decompile, or extract code from our Services; . . . (e) create accounts for our Services through unauthorized or automated means; . . . (i) create software or APIs that function substantially the same as our Services and offer them for use by third parties in an unauthorized manner.”).

²⁴² E.g., Tulir Asokan, *Whatsmeow*, GITHUB, <https://github.com/tulir/whatsmeow/> (last visited July 13, 2023) [<https://perma.cc/NZS5-9NV2>]; Sigalor, *WhatsApp Web Reverse Engineered*, GITHUB, <https://github.com/sigalor/whatsapp-web-reveng> (last visited July 13, 2023) [<https://perma.cc/GM2V-Q46F>].

WhatsApp may not be able to enforce these restrictions as a side effect of its decision to use end-to-end encryption. Since WhatsApp messages are surrounded by a layer of encryption, what is inside the layer of encryption is incomprehensible to anyone except a message’s recipient. It may be that WhatsApp has no way of forcing any app from using the *forwardingScore* metadata.

In fact, our research conclusively confirms that the source code of some third-party clients bypass *forwardingScore* altogether. We have found source code repositories for WhatsApp clients that ignore this entire functionality, not even displaying messages as “Forwarded” / “Forwarded many times.” This strongly suggests that when WhatsApp rolled out the forwarding score limits in 2020, it was unable to force all the third-party clients to do so as well.

Third-party apps are in significant use. One app that ignores the forwarding limits, “whatsmeow,” is one of the top five percent packages on a prominent list of the most popular software packages written in the “Go” programming language.²⁴³ A WhatsApp third-party client community had over 700,000 users in 2015.²⁴⁴ The aforementioned bot “Bosco,” which also ignores the restrictions, has been “forked” over 1,200 times, and we even found a YouTube tutorial on how to deploy it.²⁴⁵

V. GUIDELINES FOR POLICYMAKERS MANDATING FRICTION

As described in detail in Part II, false news is a grave concern for democracies around the world. There is growing debate in democratic nations on the measures that should legitimately be adopted to reduce the spread of false news, limit its impact on individuals and promote

²⁴³ *Github.com/tulir/whatsmeow*, ECOSYSTEMS: PACKAGES, <https://packages.ecosyste.ms/registries/proxy.golang.org/packages/github.com%2Ftulir%2Fwhatsmeow> (last updated June 29, 2023) [<https://perma.cc/DP4Y-DCR4>] (“Top 8.2% on proxy.golang.org.”).

²⁴⁴ Stephen Hall, *The Most Popular 3rd-Party WhatsApp Client Is Now Dead*, 9TO5GOOGLE (Jan. 21, 2015, 9:26 AM PST), <https://9to5google.com/2015/01/21/whatsapp-plus-cease-and-desist/> [<https://perma.cc/RF6U-SF36>].

²⁴⁵ Pepesir, *supra* note 239; PEPE SIR, *DEPLOY FULL BUTTON WHATSAPP BOT IN HEROKU* 📧 ⚡ 24 HOURS ONLINE 📧 NO TERMUX 📧 📧 WORKING|BOSCO BOT 📧 NO ENC 📧, YOUTUBE (Dec. 4, 2021), <https://www.youtube.com/watch?v=ZJQ5owYh7dc> [<https://perma.cc/TBS5-VXL7>].

meaningful public discourse based on true, honest, and impartial information.²⁴⁶ Allowing false information to spread widely undermines people's ability to generate an informed opinion on meaningful issues under public debate. Elections that are based on false information undermine the very essence of democracy. As Hans Kelsen described: "The will of the community, in a democracy, is always created through a running discussion between majority and minority, *through free consideration* of arguments for and against a certain regulation of a subject matter."²⁴⁷

We suggest that friction can be adopted as a meaningful tool in limiting the spread and overcoming the harms of false news, despite our ability to find ways to circumvent WhatsApp's implementation. We recommend adopting friction as a complementary tool to other efforts such as investing in digital literacy, encouraging platforms to work together with reliable fact checkers, reducing financial incentives for actors benefitting from false news, improving online accountability, and other initiatives.

Our deep dive into the infrastructure involved in the WhatsApp forwarding friction provides important lessons for policymakers considering enacting legislation or promulgating rules mandating friction in the design of technology platforms. They point to recurring considerations that will appear in (almost) any case where friction is used and must be taken into account when thinking about the type of challenges friction can overcome, as well as how to design and implement it.

²⁴⁶ See, e.g., Andreas Jungherr & Ralph Schroeder, *Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy*, 7 Soc. Media & Soc'y, Jan.-Mar. 2021, at 1 (arguing that regulation should focus on the structural transformations facilitating the spread of data); Morgan, *supra* note 2, at 41-43 (describing initiatives by tech companies to limit election manipulation); Niels Nagelhus Schia & Lars Gjesvik, *Hacking Democracy: Managing Influence Campaigns and Disinformation in the Digital Age*, 5 J. Cyber Pol'y 413, 417-23 (2020) (comparing the way that Norway and the U.K. have dealt with attempts to use technology to subvert democratic processes); Chris Tenove, *Protecting Democracy from Disinformation: Normative Threats and Policy Responses*, 25 Int'l J. Press/Pol. 517, 519-21 (2020) (proposing that policies addressing misinformation are actually protecting self-determination, accountable representation, and public deliberation).

²⁴⁷ HANS KELSEN, GENERAL THEORY OF LAW AND STATE 287-288 (Anders Wedberg trans., Russell & Russell 1961).

Those seeking to mandate or implement friction must consider four central guidelines. The *first* is based on the decades-old understanding that code is law.²⁴⁸ The technical analysis we conducted teaches us the importance of understanding code for policymakers. Law regulating the activity of technology companies must be grounded in a deep understanding of the technical aspects of the regulated activity.

The *second* insight is concerned with the content-neutral nature of WhatsApp's restrictions. A central challenge democracies face when seeking to limit the spread of false news is the desire to simultaneously protect the freedom of expression, such as potential challenges brought under the First Amendment. Any content-based restriction on the spread of speech will likely be found unconstitutional. Frischmann and Benesch have argued that the way in which WhatsApp implemented friction in its end-to-end encrypted setting is content-neutral.²⁴⁹ Our technical analysis confirms that WhatsApp has implemented these features in a content-neutral manner.

Our *third* guideline calls on policymakers to consider how effective the integration of friction is likely to be. As we learned in Part II.B, false news spreads wider and faster than true news. Part of the reason for this is the ability to purposely design false news to be emotionally evocative.²⁵⁰ It is sensationalist, inflammatory, dramatic, and shocking. Thus, people are more motivated to spread it. In order to ensure that the friction does not defeat its purpose by limiting the spread of true news while allowing false news to continue spreading like wildfire, it is important for policymakers to have a tool to test the actual effect of the friction. Measurement will also be important regarding the ability to determine the different effects that friction has on different actors. Thus, highly motivated superusers may be susceptible to friction that the average user is impacted by.

Fourth, and finally, we observe that the introduction of friction is merely the beginning of the process, not its end. Based on its effectiveness, friction can be turned up or down. Policymakers may have an interest in changing its level (for example, increasing friction in the

²⁴⁸ LESSIG, *supra* note 7, at 5.

²⁴⁹ Frischmann & Benesch, *supra* note 9, at 441.

²⁵⁰ Pennycook & Rand, *supra* note 12, at 393.

months before elections). Even if this is not the case, the introduction of friction will likely ignite an arms race, encouraging motivated superusers of the friction to find ways around it. In this respect, as well as others, friction can be an effective intermediary tool to limit the spread of false news. Friction can provide regulators with the time they need to continue thinking about effective ways to limit the spread and tackle the harms caused by false news.

A. *Code Is Law; Law Should Be (Based on an Understanding of) Code*

This article joins a movement in legal scholarship focused on ways to harness the power of design as a means for addressing harms wrought by technological change. The movement builds on insights from Science and Technology Studies (“STS”) that interrogate the ways in which the design of technological artifacts advance or impede particular human values.²⁵¹

Legal scholars have built on this STS foundation to suggest how law and legal institutions can play a key role in encouraging value-sensitive design.²⁵² Woodrow Hartzog encourages legislators and regulators to nudge technology companies to design online services that better respect user privacy.²⁵³ Ari Waldman focuses on the mechanisms that derail the design goals of privacy-minded professionals working for these companies.²⁵⁴

²⁵¹ See Langdon Winner, *Do Artifacts Have Politics?*, 109 DAEDALUS 121, 122-23 (1980); see also, e.g., Mary Flanagan, Daniel C. Howe & Helen Nissenbaum, *Embodying Values in Technology: Theory and Practice*, in INFORMATION TECHNOLOGY AND MORAL PHILOSOPHY 322, 331-47 (Jeroen van den Hoven & John Weckert eds., 2008) (using design of a video game as a case study for embedding values into technological systems); BATYA FRIEDMAN & DAVID G. HENDRY, VALUE SENSITIVE DESIGN: SHAPING TECHNOLOGY WITH MORAL IMAGINATION 16 (2019).

²⁵² WOODROW HARTZOG, PRIVACY’S BLUEPRINT 157-94 (2018); WALDMAN, *supra* note 43, at 232-49. Frischmann and Selinger examine one application of this principle. They highlight the problematic nature of consumers almost automatically consenting to online contracts they are presented with. To ensure that users conduct even a minimal deliberative process, they suggest combining “meaningful notice with speed bumps [in our terminology – friction-in-design] to prompt demonstrable deliberation.” FRISCHMANN & SELINGER, *supra* note 44, at 291.

²⁵³ HARTZOG, *supra* note 252, at 157-94.

²⁵⁴ WALDMAN, *supra* note 43, at 232-49.

Friction belongs at the center of this work. It is a pervasive yet underappreciated design pattern that can give rise to the time and space necessary to protect human values, including privacy, trust, and safety.²⁵⁵

Treating friction and design as a subject of regulatory attention helps realize the unrealized promise of Reidenberg's and Lessig's focus on code is law.²⁵⁶ For too long, legal scholars have focused solely on the descriptive version of this maxim: code operates like law, so we ought to study the way code constrains human behavior alongside law.²⁵⁷ Our study of friction deploys a more proactive, less studied, prescriptive version of the maxim: we should focus on enacting laws that nudge, shape, harness, or ban particular forms of code to combat false news and other information age harms.²⁵⁸

A key predicate for harnessing the prescriptive dimension of "code is law" is interdisciplinary engagement. Legal scholars and technology scholars must collaborate to credibly build out the design agenda. Our interdisciplinary investigation of WhatsApp, taking advantage of our mixed legal and technical training, was necessary to explicate exactly how WhatsApp had tried to harness friction to fight false news. We came to understand the many limitations, tradeoffs, and shortcuts WhatsApp had taken, helping us unearth the lessons for regulators who want to harness this kind of friction themselves.

B. Content-Neutrality

Democracies are more susceptible to false news than other forms of government. One of the reasons for this is the fundamental respect they have for free speech.²⁵⁹ In the United States the First Amendment prevents the government from mandating regulation that abridges

²⁵⁵ See generally CHRISTOPHER ALEXANDER, SARA ISHIKAWA, MURRAY SILVERSTEIN, MAX JACOBSON, INGRID FIKSDAHL-KING & SHLOMO ANGEL, *A PATTERN LANGUAGE: TOWNS, BUILDINGS, CONSTRUCTION* (1977) (explaining how pattern language can solve design problems).

²⁵⁶ LESSIG, *supra* note 7, at 1-8; Reidenberg, *supra* note 7, at 554-55.

²⁵⁷ Paul Ohm, *Computer Programming and the Law: A New Research Agenda*, 54 VILL. L. REV. 117, 144 (2009).

²⁵⁸ Ohm, *supra* note 26, at 1373-74.

²⁵⁹ See Cohen, *supra* note 95, at 659; McKay & Tenove, *supra* note 98, at 706.

individuals' freedom of speech.²⁶⁰ While the government may not restrict speech based on its content, it is allowed to create reasonable restrictions on the time, place, and manner ("TPM") of speech, as long as they are content-neutral.²⁶¹ Therefore, the question of content-neutrality has been labeled "the central inquiry" in TPM restrictions.²⁶² In protecting this basic constitutional right, the United States Supreme Court ruled that "government has no power to restrict expression because of its message, its ideas, its subject matter, or its content."²⁶³

Content-neutral restrictions may be found justified as long as they are "narrowly tailored to serve significant governmental interests, and that [they] leave open ample alternative channels for the communication of the information."²⁶⁴ Deciding whether a restriction is content-based has meaningful judicial implications. While content-based restrictions must face strict judicial scrutiny under the First Amendment, content-neutral restrictions face intermediate scrutiny, a lower judicial standard.²⁶⁵

One of the prime advantages of friction as a tool to limit the spread of false news, compared to other options, is its strong content-neutral nature. In their recent article, Frischmann and Benesch explain that many cases of friction can actually be viewed as content-neutral TPM

²⁶⁰ U.S. CONST. amend. I.

²⁶¹ *Cox v. New Hampshire*, 312 U.S. 569, 576 (1941).

²⁶² Erwin Chemerinsky, *Content Neutrality as a Central Problem of Freedom of Speech: Problems in the Supreme Court's Application*, 74 S. CAL. L. REV. 49, 49 (2000).

²⁶³ *Police Dep't v. Mosley*, 408 U.S. 92, 95 (1972).

²⁶⁴ *Ward v. Rock Against Racism*, 491 U.S. 781, 791 (1989).

²⁶⁵ See Frischmann & Benesch, *supra* note 9, at 424-25. The Supreme Court first applied the TPM restriction in a case involving a group of Jehovah's Witnesses who organized a march without obtaining the necessary authorization. The Court held that the law requiring them to obtain prior authorization was constitutional as long as it was not applied in a discriminatory (i.e. content-based) fashion. *Ward*, 491 U.S. at 791. Justice Hughes recognized in his opinion that an organized society was a prerequisite for the exercise of constitutional civil liberties. With "too much" liberty, no liberty would in fact be protected. In *Reed v. Town of Gilbert*, 576 U.S. 155 (2015), the Supreme Court clarified that in determining whether a restriction was content based or not, the government's purpose in the restriction was not the central inquiry. Justice Thomas ruled that a restriction will not be considered content-neutral if it draws a distinction based on the speaker's message. A restriction based on the content of the message spoken would be subject to strict judicial scrutiny and would likely fail it. See David L. Hudson, Jr., *The Content-Discrimination Principle and the Impact of Reed v. Town of Gilbert*, 70 CASE W. RES. L. REV. 259, 261 (2019).

restrictions.²⁶⁶ In order to be considered content-neutral, a government-mandated restriction on speech must “not treat speakers (or authors) differently according to what type of speech, sign, music etc. they wanted to disseminate.”²⁶⁷ In order to withstand judicial scrutiny such a restriction would have to fulfill three requirements. It would have to be “[1] narrowly tailored to [2] serve a substantial government interest and [3] do not unreasonably limit alternative avenues of expression.”²⁶⁸

These requirements are not easy to withstand, and many types of restrictions on speech will likely be found to be content based when faced with constitutional scrutiny. In certain cases, friction may indeed be the only tool able to constitutionally restrict speech.²⁶⁹ In other cases, government restrictions would likely fail. One such example is content moderation on social media platforms.²⁷⁰ Social media platforms engage in content moderation all the time.²⁷¹ They do this because providing users with complete, unrestricted freedom of expression would cause their platforms to be overrun with bullying, hate speech, and pornography.²⁷² Each social media platform has their own community guidelines on what types of speech are permitted, and what types will be blocked (either by AI or by humans).²⁷³ The government cannot do the same.²⁷⁴ It cannot create content-based restrictions on speech.²⁷⁵ At the

²⁶⁶ Frischmann & Benesch, *supra* note 9, at 44.

²⁶⁷ *Id.*, *supra* note 9, at 424.

²⁶⁸ *Ward*, 491 U.S. at 789.

²⁶⁹ See Frischmann & Benesch, *supra* note 9, at 382 (detailing the restrictions of current regulatory approaches).

²⁷⁰ See Douek, *supra* note 3, at 26.

²⁷¹ See Klonick, *The New Governors*, *supra* note 3, at 1611, 1659.

²⁷² See, Klonick, *The New Governors*, *supra* note 3, at 1640.

²⁷³ E.g., *Facebook Community Standards*, FACEBOOK TRANSPARENCY CTR., <https://transparency.fb.com/policies/community-standards/> (last visited July 8, 2023) [<https://perma.cc/P5E6-34PE>]; *Community Guidelines*, INSTAGRAM HELP CTR., <https://help.instagram.com/477434105621119> (last visited July 8, 2023) [<https://perma.cc/HZ2H-958H>]; *Community Guidelines*, YOUTUBE, <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/> (last visited July 8, 2023) [<https://perma.cc/V62Z-VJUJ>]; *The Twitter Rules*, TWITTER HELP CTR., <https://help.twitter.com/en/rules-and-policies/twitter-rules> (last visited July 8, 2023) [<https://perma.cc/389C-ZPTU>].

²⁷⁴ See U.S. CONST. amend. I.

²⁷⁵ See *id.*

same time, the government has a legitimate interest in protecting democracy against false news. More specifically democratic governments have a deep interest in ensuring that elections are conducted in an environment free from manipulation by adversaries. In such cases friction may very well be the last resort of democratic governments. Prevented from restricting false news based on its content, content-neutral friction can be governments' last resort in their attempts to restrict the free flow of false news.

Regulators planning on introducing friction as means to restrict speech who want to ensure that their restriction is completely content-neutral can draw inspiration from the WhatsApp example. Content-neutrality is neither necessary nor sufficient with friction, but many friction-based approaches will turn out to be content-neutral.

C. Does It Really Work?

The implementation of friction by WhatsApp was done to limit the spread of false news. In order to assess its effectiveness one would need to analyze its results and outcomes. Has the intervention been effective? This question requires examining two perspectives. First, by focusing on the content. How many false positives has the friction generated versus false negatives? In other words, how many true pieces of content has the friction slowed down compared to how many articles of false news have still been spread far and wide? For friction targeted at restricting the spread of false news in an end-to-end encrypted environment like WhatsApp, the question of measurement remains particularly challenging.

1. On the Importance of Being Measured

Since imposing restrictions on forwarding, WhatsApp reported a seventy percent drop in the spread of "highly forwarded" content.²⁷⁶ As the limitation on forwarding messages is content-neutral and because of the end-to-end encryption, WhatsApp does not know the nature of the content being spread less due to its decision.²⁷⁷ Given the

²⁷⁶ Porter, *supra* note 188; Spring, *supra* note 188.

²⁷⁷ See *Keeping WhatsApp Personal and Private*, *supra* note 186.

sensationalist nature of false news and users' motivations for responding strongly to such content (or at least to the title of the content) there may be valid room for concern regarding the effect of the friction. If it turns out that false news is so much more attractive for sharing than real content that it continues spreading despite the friction, this should cause us to question the effectiveness of the type of friction introduced. If, however, the drop of seventy percent in the spread of content can be attributed mostly to a decline in the spread of false news, the intervention can be crowned a success. Given the encrypted nature even WhatsApp cannot know for certain what content is restricted by the friction. Without the ability to design a mechanism allowing us a clear answer, it will be necessary for policymakers to make a decision about the integration of such tools without full information on the distribution of content it impacts.

The ability to collect data about the exact effect of friction on different types of content (for example in carefully tailored experiments) is not enough. The next step requires policymakers to decide what is considered an acceptable trade off. How many pieces of true news and other types of neutral content are we willing to have compromised per each unit of false news whose spread is limited by the friction? If we discovered that the introduction of friction lowered the spread of false news by only ten percent and the sharing of cat videos by ninety percent, would that be considered an acceptable trade off?

2. What Type of User Is Impacted

Measuring the impact of friction to assess its effectiveness is not limited to the type of content whose spread is inhibited by the friction. Different users may be impacted differently by the introduction of friction as well. Over a decade ago, Paul Ohm published *The Myth of the Superuser: Fear, Risk, and Harm Online*.²⁷⁸ In it, Ohm describes a systemic bias influencing policymakers when thinking about internet users. When designing policy, these actors have a particular type of user in mind. Dubbed the superuser, “[h]e (always he) is a mythical figure: difficult to find, expensive to catch, able to circumvent any

²⁷⁸ Ohm, *supra* note 26.

technological constraint, and aware of every legal loophole.”²⁷⁹ In this respect, the creators of disinformation may be considered superusers. They are highly motivated, organized actors using the tool of disinformation to promote ideological goals. They may be incentivized to invest time and effort to promote their false content. Most users, however, are far from this depiction. Instead, they are “unsophisticated, exercising limited power and finding themselves restricted by technological constraints.”²⁸⁰ While acknowledging that superusers do exist and are a concern for internet users, Ohm criticizes policymakers for designing policy with the superuser in mind, “resulting in overbroad prohibitions, harms to civil liberties, wasted law enforcement resources, and misallocated economic investment.”²⁸¹

The introduction of friction has a different impact on regular users as compared to superusers (and thus, on the spread of *misinformation* compared to that of *disinformation* respectively). Our analysis in Part IV uncovered that it is possible to circumvent the friction introduced by WhatsApp without the need for a very high level of technical expertise. Users with even moderate technical skill can set the source code breakpoint and type the short lines of code we discovered to circumvent all of WhatsApp’s restrictions. Even nontechnical users can download a plug-in or install a bot to bypass the restrictions, and our research found numerous examples of such plug-ins and bots in the wild. Actors spreading disinformation may be motivated to search for and exploit these vulnerabilities. Friction will, however, restrict the downstream spread of such content by unsuspecting, unprofessional users.

Ohm recognizes that in some contexts, superusers are prevalent and worthy of regulatory attention.²⁸² In the context of false news, particularly in the period leading up to elections, superusers are far from a myth. While the challenge of foreign electoral intervention in elections has existed for decades,²⁸³ social media has exacerbated the risk for such

²⁷⁹ *Id.* at 1330.

²⁸⁰ *Id.*

²⁸¹ *Id.* at 1328.

²⁸² *Id.* at 1396-97.

²⁸³ See Jonathan J. Godinez, *The Vested Interest Theory: Novel Methodology Examining US-Foreign Electoral Intervention*, 11 *J. Strategic Sec.*, 2018, at 1; Dov H. Levin, *Partisan Electoral Interventions by the Great Powers: Introducing the PEIG Dataset*, 36 *Conflict*

intervention as well as their effectiveness.²⁸⁴ Countries now use social media to covertly intervene in the political and electoral processes of other countries. These are well-funded and technically sophisticated actors; in short, they are superusers. Russia had strong motives and a host of capabilities and opportunities to intervene in the American electoral process, as had been demonstrated and uncovered during the period leading up to the 2016 U.S. Presidential election.²⁸⁵ Russia, other countries, large organizations, and motivated political candidates *are* all superusers. There is no way for us to definitively determine how many superusers have already found the shortcomings in the technical implementation of the restrictions on forwarding that we found, and perhaps others. In Part IV we established that many different third-party apps, with thousands of users, have circumvented WhatsApp's forwarding limits and marking of "forwarded" and "frequently forwarded" messages. Since solo developers working on open-source projects have found these workarounds, this is strong evidence that a motivated, sophisticated actor like Russia could have, and probably has, done the same and may be using this vulnerability to spread disinformation. In the context of code that is relatively easy to manipulate, as we have uncovered is the case with the WhatsApp restrictions on forwarding, the superuser is a force to be reckoned with.

Mgmt. & Peace Sci. 88, 88 (2016); Vasu Mohan & Alan Wall, *Foreign Electoral Interference: Past, Present and Future*, 20 Geo. J. Int'l Affs. 110, 110 (2019).

²⁸⁴ See 5 SELECT COMM. ON INTEL., U.S. SENATE, 116TH CONG., REPORT ON RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION: COUNTERINTELLIGENCE THREATS AND VULNERABILITIES 171, 176, 183, 204 (2020) ("The GRU [a Russian intelligence military agency] also relied on U.S. social media platforms and media attention for its influence operations." One of Russia's methods of operation was by "the use of fake personas on social media." "The GRU used . . . social media personas to promote and disseminate stolen documents from the DNC and DCCC, while obscuring the GRU's involvement in the influence campaign." "Throughout the 2016 U.S. presidential campaign, Russia Today ("RT") and Sputnik used their social media accounts to push WikiLeaks-related content that disparaged Hillary Clinton. On at least two occasions, RT announced WikiLeaks releases on Twitter prior to WikiLeaks making that announcement itself").

²⁸⁵ MUELLER, *supra* note 106, at 14; see Kevin Collier, *Researchers Discover Sprawling Pro-U.S. Social Media Influence Campaign*, NBC NEWS (Aug. 24, 2022, 1:55 PM PST), <https://www.nbcnews.com/tech/misinformation/researchers-discover-sprawling-us-social-media-influence-campaign-rcna44595> [<https://perma.cc/S8GE-4677>].

D. *It's a Never-Ending Story*

The integration of friction into technology does not end with a single decision. Instead, reaching the right level of friction requires tuning the level of friction based on the goal the friction is aimed at achieving as well as on its effects de facto. In this Section we detail the various considerations involved in reaching the “right” level of friction.

1. Tunability

Unlike other forms of regulation, imposing friction is not simply a question of permission versus prohibition. The introduction of friction generates a certain measure of technical inefficiency in order to achieve a desired goal.²⁸⁶ The exact level of inefficiency necessary to achieve the goal will usually not be predetermined or even clear.²⁸⁷ As presented in Part III, the friction introduced by WhatsApp into its forwarding mechanism has changed over time. Finding the balance between not enough friction and too much friction requires delicate experimentation.²⁸⁸ Unlike much of command-and-control regulation which is binary (the regulation applies, or does not apply), friction is not an all-or-nothing choice.²⁸⁹ It is possible and desirable to “choose to calibrate the amount of friction at an infinite number of levels, responsive to the costs and benefits in each situation.”²⁹⁰ The level of friction can be set and readjusted as regulators or software engineers examine its effects and outcomes.²⁹¹ The ease with which friction can be tuned is one of its advantages. Tuning and retuning software is a familiar process for platforms. Technology companies routinely test various aspects of software and services, updating them as necessary.²⁹²

²⁸⁶ Ohm & Frankle, *supra* note 5, at 783.

²⁸⁷ *Id.* at 820.

²⁸⁸ Frischmann & Benesch, *supra* note 9, at 389 (“We don’t want . . . too much [friction, either:] [i]t is, after all, costly. . . . Paralysis by analysis is a frightening prospect.”).

²⁸⁹ See Ohm & Frankle, *supra* note 5, at 830-31. When a company decides that after six attempts to unlock a password the phone locks, why is this okay? Why should it not be eight? Or three? See *id.* at 816.

²⁹⁰ McGeeveran, *supra* note 4, at 53.

²⁹¹ See Ohm & Frankle, *supra* note 5, at 819.

²⁹² McGeeveran, *supra* note 4, at 61.

This feature is not unique to tools using friction and is therefore familiar to technology companies.²⁹³ They can tune, test, and retune until they reach the right balance between too much friction and too little of it. Companies famously engage in constant tuning by running A/B experiments in which companies test users' responses to "variations in page characteristics from layout to buttons to fonts."²⁹⁴

In the context of speech, the decision at which level to set the friction is not merely a technical one made by the technology companies. The level of friction cannot be based solely on the desired outcome as set by the platforms. Inasmuch as government-mandated friction restricts speech, it is important to ensure that the friction is narrowly tailored so it can withstand judicial scrutiny.²⁹⁵

The tuning of friction does not only have to move in one direction. In the context of false news there very well may be reasons to raise and lower the level of friction over time and in response to current events. In the period leading up to an election, it may be reasonable to implement a higher level of friction. In this context it will be even more important to ensure that the friction facilitates free and democratic elections and does not serve as a way to limit speech. If a messaging system or social media platform detects the organized spread of disinformation, an increase in friction may be the appropriate response.

2. Friction Will Ignite an Arms Race

Friction introduced into technology cannot be expected to stay effective over time. This is both because of technology and because of

²⁹³ See Andrea Arcuri & Gordon Fraser, *Parameter Tuning or Default Values? An Empirical Investigation in Search-Based Software Engineering*, 18 *EMPIRICAL SOFTWARE ENG'G* 594, 594 (2013) (describing the growth and importance of tuning of parameters in search-based software engineering).

²⁹⁴ ZUBOFF, *supra* note 35, at 298 (pointing to A/B testing as an integral part of what she calls surveillance capitalism); see Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle & James H. Fowler, *A 61-Million-Person Experiment in Social Influence and Political Mobilization*, 489 *NATURE* 295, 295 (2012) (describing an A/B-based experiment on Facebook to test how subtle variations in messages on the platform could influence political mobilization).

²⁹⁵ In *Ward v. Rock Against Racism*, 491 U.S. 781, 798 (1989), the Supreme Court held that "[the] regulation of the time, place or manner of protected speech . . . need not be the least restrictive or least intrusive means of doing so."

the human factor. First, friction that was introduced at a particular time in the development of technology may become obsolete as it rapidly develops. New types of friction will have to be developed and implemented. Second, almost as soon as a technological barrier is put up, parties with a large stake in the technology will start trying to circumvent the limitations imposed by friction.²⁹⁶ In particular, superusers, discussed above, can be expected to invest resources into finding ways to bypass or lower the level of friction imposed by the technology company. In the context of WhatsApp's limits on forwarding, we found that circumventing the platform's restrictions is relatively easy. We also found numerous cases in the wild of developers writing and deploying code that circumvents the forwarding restrictions.²⁹⁷

One example of an arms race can be found in the use of bots by parties with an interest in spreading their message broadly. Leading platforms have made various attempts at identifying bots active on them and banning them from the site.²⁹⁸ This does not stop new ones from being developed and deployed on the very same platforms, which will then develop new tools to identify and remove them. Researchers explain, "bot detection will always be a cat-and-mouse game in which a large, but

²⁹⁶ See Kovarsky, *supra* note 27, at 933.

²⁹⁷ See *supra* Part IV.E.2.

²⁹⁸ See Craig Timberg & Elizabeth Dwoskin, *Twitter Is Sweeping out Fake Accounts Like Never Before, Putting User Growth at Risk*, WASH. POST (July 6, 2018, 6:30 PM EST), <https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/> [https://perma.cc/J7VS-ARW4]; Christopher Bing, *Exclusive: Twitter Deletes over 10,000 Accounts That Sought to Discourage U.S. Voting*, REUTERS (Nov. 2, 2018, 1:03 PM), <https://www.reuters.com/article/us-usa-election-twitter-exclusive/exclusive-twitter-deletes-over-10000-accounts-that-sought-to-discour3ge-u-s-voting-idUSKCN1N72FA> [https://perma.cc/74ZB-UPYB]; Andrew Hutchinson, *Instagram Looks to Crackdown on Bots with New Review and ID Process*, SOC. MEDIA TODAY (Aug. 13, 2020), <https://www.socialmediatoday.com/news/instagram-looks-to-crackdown-on-bots-with-new-review-and-id-process/583491/> [https://perma.cc/4NKP-2RA3]; Jay Peters, *YouTube's New Weapons for Fighting Comment Spam Include 24-Hour Bans*, VERGE (Dec. 13, 2022, 2:35 PM PST), <https://www.theverge.com/2022/12/13/23508062/youtube-comment-spam-warning-violation-detection-bots> [https://perma.cc/Y9WR-4EBX]; Mariel Soto Reyes, *Facebook Removes 2.2 Billion Fake Accounts in Three Months*, BUS. INSIDER (May 28, 2019, 8:31 AM PDT), <https://www.businessinsider.com/facebook-removed-22-billion-fake-accounts-2019-5> [https://perma.cc/47HV-97MC].

unknown, number of humanlike bots may go undetected. Any success at detection, in turn, will inspire future countermeasures by bot producers.”²⁹⁹ Regulators will mandate friction, platforms will implement it in one way or another, interested users will try to circumvent it, platforms will develop an updated version of the friction and so on and so forth. Any technical limitation will lead to an ongoing arms race with motivated actors spreading disinformation. At the same time, regular users spreading misinformation are unlikely to be engaging in such an arms race.

E. Friction Can Be Used as a Temporary Fix

The examples of friction we discussed in this Article have at least one thing in common. They address but do not solve the underlying problem of the prevalence of false news. Friction can sometimes be a good intermediate solution without necessarily being the end goal. Friction can provide policymakers with time they may need to address the underlying cause of the problem. In the case of limiting the spread of false news, steps like those adopted by WhatsApp might not remain effective over time. Highly motivated users will find ways to work around the technology used to implement the friction. The very same technology will become obsolete as new technology rapidly develops. During this time afforded by friction, policymakers should invest in efforts in more directly attacking the root causes of false news and the harms generated by it.

Regulation mandating friction must take this temporal effect into account. This can be done using several regulatory techniques. First, regulation mandating friction can use broad standards (for example defining the goal friction should aspire to). The rules themselves (what type of friction, where in the technology it should be integrated, what level it should be tuned to) can be detailed by regulators such as the Federal Election Commission or Federal Trade Commission, depending on the industry sector and goal. Another legislative tool that can be used to promote time-limited technology is the sunset clause. Sunset clauses are used in cases where a law or regulation requires periodic

²⁹⁹ Lazer et al., *supra* note 8, at 1095.

revisiting.³⁰⁰ Policymakers should use the timeframe during which friction acts as a temporary barrier against the broad spread of false news, to find more permanent solutions against false news.

Friction does not attack the underlying challenges of false news. It cannot make superusers less motivated to generate and spread disinformation (though it will make it more costly to spread it), nor will it change human psychology leading people to believe and share misinformation (though it does make an attempt to channel psychological insights to nudge people away from broadly sharing it).³⁰¹ Friction is a temporary cure for the symptom of false news, an important tool for keeping democracies safe and stable while policymakers grapple with the underlying reasons driving false news and design a legal foundation to hold its spreaders accountable for the broad societal damage they create.

CONCLUSION

The first thirty years of the commercial internet has seen the technology industry focus on rapid growth, dramatic change, and a single-minded search for efficiency. We might even call this the Age of Frictionlessness. Today, we hope the growing appreciation by technologists and legal scholars for the power and potential of friction to attack wicked problems like false news marks the dawn of a new age in which friction and frictionlessness are seen as complementary forces.

By studying the WhatsApp forwarding example closely, this Article illuminates some of the benefits that friction brings to those who deploy it. Friction lends itself to *tunable* solutions, giving regulators a rheostatic feedback level of control over the strength of their solutions. Turn the friction dial up (cut the number of chats to which a message may be forwarded) and reduce the spread of both false news and valuable speech; turn the friction dial down (increase the number of people in a group) and increase the spread of both. This provides a more supple, flexible approach to the regulation of technology than other, more conventional proposals.

³⁰⁰ See Sofia Ranchordás, *Innovation-Friendly Regulation: The Sunset of Regulation, The Sunrise of Innovation*, 55 JURIMETRICS J. 201, 219 (2015).

³⁰¹ See Calo et al., *supra* note 109, at 3.

Friction wielders must also understand that they are entering into inevitable *arms races*. Highly skilled and motivated actors will likely find ways to circumvent various types of friction, as we did with WhatsApp's restrictions. Regulators and technology companies must decide which circumventions are worth preventing and which to let slip through, depending on whether they are more worried about superusers, regular users, or both.

Although we offer these guidelines, we also raise new questions and hope this Article sets an agenda for future research in this space. Our study did not test WhatsApp's claims that their friction had reduced virality by seventy percent.³⁰² Calculating this statistic requires a network-level view of activity that our techniques alone cannot measure. This research will be difficult to undertake, as the end-to-end encryption protecting WhatsApp users from government surveillance also blocks the view of researchers, so creative new approaches are needed to test the efficacy of this approach. Moreover, knowing that *virality* has reduced seventy percent does not reveal the division between false news and benign memes that have been affected.

Many questions remain, but the overall agenda of integrating friction into technology is extremely promising. Friction has the potential to provide policymakers with a new, tunable, content-neutral set of tools to allow them to address challenges that would otherwise be deemed intractable.

³⁰² Spring, *supra* note 188, at 1.