Northwestern Journal of Human Rights

Volume 21
Issue 2 *Spring*

Article 1

Spring 3-2023

Managing Misinformation on Social Media: Targeted Newsfeed Interventions and Freedom of Thought

Richard Mackenzie-Gray Scott
richard.mackenzie-grayscott@law.ox.ac.uk

Managing Misinformation on Social Media: Targeted Newsfeed Interventions and Freedom of Thought

Cover Page Footnote

Postdoctoral Fellow, Bonavero Institute of Human Rights, and Fellow, St Antony's College, University of Oxford: richard.mackenzie-grayscott@law.ox.ac.uk. Research funded by the British Academy (grant no. BAR00550-BA00.01). With many thanks to Aradhana Cherupara Vadekkethil, Damian Tambini, Daniele Nunes, Elena Abrusci, Geoffrey Chapman, John Croker, Kate O'Regan, Katie Pentney, Keri Grieman, Martin Scheinin, Nerissa Naidoo, Oliver Butler, Rasmus Kleis Nielsen, Rossella Pulvirenti, Six Silberman, Talita Dias, Tanya Krupiy, and Tarik Gherbaoui for their helpful comments on the original draft and/or contributions to the workshop Digital Realm Challenges to Freedom of Expression, held at the Bonavero Institute on 21 April 2022, where this paper was first presented. Many thanks also to Reuben Binns for the opportunity to present this research to him and his colleagues in the Human Centered Computing Group at the Department of Computer Science, University of Oxford on 8 November 2022, where the discussion and feedback was constructive and interesting. With thanks also to the editors and reviewers at the Journal for their guidance and support, in particular Zoë Reinstein, Jordan Plummer, Samantha Reilly, and Taylor Desgrosseilliers.

Volume 21, Number 2 (2023)
Northwestern Journal of Human Rights

MANAGING MISINFORMATION ON SOCIAL MEDIA: TARGETED NEWSFEED INTERVENTIONS AND FREEDOM OF THOUGHT

Richard Mackenzie-Gray Scott[*]

ABSTRACT—Whether it is being told a particular politician consumes children, or drinking cow urine will cure your disease, or that Jimi Hendrix is alive and well living the good life in Drumnadrochit, misinformation affects societies in myriad ways. Its spread online via social media platforms raises questions concerning how it can be addressed. This article engages with a related problem: Can the use of targeted behavioral interventions on social media newsfeeds to reduce the spread of misinformation be reconciled with the human right to freedom of thought?

"*Lock up your libraries if you like; but there is no gate, no lock, no bolt that you can set upon the freedom of my mind.*"
Virginia Woolf

"*Hold my beer*."
Zuckerberg et. al.

[*] Postdoctoral Fellow, Bonavero Institute of Human Rights, and Fellow, St Antony's College, University of Oxford: richard.mackenzie-grayscott@law.ox.ac.uk. Research funded by the British Academy (grant no. BAR00550-BA00.01). With many thanks to Aradhana Cherupara Vadekkethil, Damian Tambini, Daniele Nunes, Elena Abrusci, Geoffrey Chapman, John Croker, Kate O'Regan, Katie Pentney, Keri Grieman, Martin Scheinin, Nerissa Naidoo, Oliver Butler, Rasmus Kleis Nielsen, Rossella Pulvirenti, Six Silberman, Talita Dias, Tanya Krupiy, and Tarik Gherbaoui for their helpful comments on the original draft and/or contributions to the workshop *Digital Realm Challenges to Freedom of Expression*, held at the Bonavero Institute on 21 April 2022, where this paper was first presented. Many thanks also to Reuben Binns for the opportunity to present this research to him and his colleagues in the Human Centered Computing Group at the Department of Computer Science, University of Oxford on 8 November 2022, where the discussion and feedback was constructive and interesting. With thanks also to the editors and reviewers at the Journal for their guidance and support, in particular Zoë Reinstein, Jordan Plummer, Samantha Reilly, and Taylor Desgrosseilliers.

TABLE OF CONTENTS

INTRODUCTION

Political polarization, distrust in government, and the prolongation of public health emergencies are a few of the problems to which misinformation contributes. Concoctions of contributory factors bring about the spread of misinformation today. Some have an established history dating back centuries; others are more recent, brought about by the boom of technological development. Since the arrival of social media into the lives of billions, misinformation can spread in the blink of an eye. This article explores potential reasons why the spread of misinformation on social media occurs and provides suggestions as to how misinformation might be addressed on these platforms.

The primary question guiding the analysis is whether the use of targeted behavioral interventions on social media newsfeeds to reduce the spread of misinformation can be reconciled with the human right to freedom of thought. Reconciliation is possible, but it raises challenges calling for careful treatment and nuance. Deliberately trying to change what people perceive, think, feel, and then do appears to be at tension with the human right to form thoughts freely and make decisions based on those thoughts. If targeted behavioral interventions on social media platforms are more commonly implemented to manage misinformation, then whether such measures are compatible with the human right to freedom of thought can guide questions on whether they are lawful and if they should be used.

In weighing the potential benefits of individually tailored user engagement against the costs from a human rights perspective, Part I of this article begins by explaining the problem of misinformation and its spread, specifically on social media platforms. Part II examines the potential of implementing targeted behavioral interventions on newsfeeds to address this problem and highlights why such a practice may interfere with the human right to freedom of thought, among other concerns. Part III lays out the law on this right at the international, regional, and domestic levels in order to assess whether the practice of attempting to change choices applicable to the consumption and sharing of information on social media is compatible with current law. Part IV concludes with a reflection on the implications of these findings, offering insights regarding how to move forward in managing misinformation on social media platforms in a way that, at the very least, respects the right of human beings to think freely while exploring the digital realm.

I. MISINFORMATION AND ITS SPREAD ONLINE

Although social networks have partly migrated online, interacting with the world without the use of an internet-capable device has been a part of human history for much longer than not. False information has impacted societies through these offline networks for centuries. In 1620, Francis Bacon wrote about the falsehoods to which human beings are susceptible and how "idols of the cave, comprising the errors made because of the peculiar nature of the individual and of the habits [they get] into,"[1] can make unfit choices that obstruct understanding while contributing to harmful outcomes. Over a hundred years later, John Adams, commenting on the work of Nicolas de Caritat arguing in favor of a free press,[2] wrote, "'There has been more new error propagated by the press in the last ten years than in a hundred years before 1798.'"[3] The free circulation of information, while vital to societies, does not eradicate error even if it helps resist it. Social circles where members feed the trolls occupying those caves within each human mind have led to the development of clans, cliques, creeds, and the conflicts between them. Misinformation has shaped such practices, which now occur as part of the lives of those who choose to be used by social media or must do so out of necessity.[4]

Before explaining the problem of misinformation that spreads online through social media platforms, it is first necessary to understand what misinformation is in order to be clear about how targeted behavioral interventions on newsfeeds might help address it. Providing this understanding also helps distinguish misinformation from disinformation, malinformation, and influence operations.

A. Taxonomy of Misinformation

While possible, it is difficult to pinpoint where misinformation lies along the spectrum between factually accurate information, misrepresented information, and information with no basis in fact. In probing what constitutes misinformation it is helpful to examine what it is *not*, and how it

---

[1] Elodie Cassan, *"A New Logic": Bacon's* Novum Organum, 29 PERSPECTIVES ON SCI. 255, 266 (2021).

[2] NICOLAS DE CONDORCET, OUTLINE OF AN HISTORICAL VIEW OF THE PROGRESS OF THE HUMAN MIND (Lang & Ustick, 1796).

[3] COLLEEN A. SHEEHAN, JAMES MADISON AND THE SPIRIT OF REPUBLICAN SELF-GOVERNMENT 40 (2009).

[4] V. Welch, et al., *Interactive Social Media Interventions to Promote Health Equity: An Overview of Reviews*, 36 HEALTH PROMOTION AND CHRONIC DISEASE PREVENTION IN CANADA: RSCH., POL. AND PRACTICE 63 (2016).

relates to other concepts such as disinformation and malinformation. Lance Bennett and Steven Livingston define disinformation as:

> [I]ntentional falsehoods or distortions, often spread as news, to advance political goals such as discrediting opponents, disrupting policy debates, influencing voters, inflaming existing social conflicts, or creating a general backdrop of confusion and informational paralysis.[5]

Their previous research shows that disinformation involves the production and dissemination of deliberately inaccurate information for the purpose of deceiving an audience.[6] This practice falls within what appears to be the broader issue of "influence operations," which Alicia Wanless notes is a "phenomenon" that "still has not been well defined."[7] Considering the variety of definitions of influence operations, it is unclear where misinformation sits in relation to this concept.[8]

This does not mean misinformation cannot be defined. Irene Khan provides a general definition of misinformation, alongside disinformation and malinformation:

> "[D]isinformation" is described as false information that is knowingly shared with the intention to cause harm, 'misinformation' as the unintentional dissemination of false information and "malinformation" as genuine information shared with the intention to cause harm.[9]

Misinformation can be described in a number of ways,[10] including context-specific definitions.[11] While these definitions leave questions about

---

[5] W. Lance Bennett & Steven Livingston, *A Brief History of the Disinformation Age: Information Wars and the Decline of Institutional Authority*, *in* THE DISINFORMATION AGE: POLITICS, TECHNOLOGY, AND DISRUPTIVE COMMC'N. IN THE UNITED STATES 3 (W. Lance Bennett & Steven Livingston eds., 2020).

[6] *See* W. Lance Bennett & S. Livingston, *The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions*, 33 EUR. J. OF COMMC'N. 122 (2018).

[7] Alicia Wanless, *What's Working and What Isn't in Researching Influence Operations?*, LAWFARE (Sept. 22, 2021, 10:32 AM), https://www.lawfareblog.com/whats-working-and-what-isnt-researching-influence-operations.

[8] *See* Alicia Wanless & James Pamment, *How Do You Define a Problem Like Influence?*, 18 J. OF INFO. WARFARE 1, 6-7 (2019).

[9] Irene Khan (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), *Disinfo. & Freedom Op. & Expression*, ¶ 12, U.N. Doc. A/HRC/47/25 (Apr. 13, 2021).

[10] *See* Claire Wardle & Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*, 6 COUNCIL OF EUR. (2017); *Misinformation*, CAMBRIDGE ENG. DICTIONARY, https://dictionary.cambridge.org/us/dictionary/english/misinformation (last visited Jan. 30, 2023); *Misinformation*, OXFORD DICTIONARY MEDIA & COMMC'N., https://www.oxfordreference.com/display/10.1093/acref/9780199568758.001.0001/acref-9780199568758-e-1748 (last

whether misinformation includes misleading but not necessarily false information, they help distinguish misinformation from disinformation. It is reasonably clear that the latter includes the intent of an actor to create or spread information contrary to established facts, whereas misinformation does not.

The significance of categorizing unintentionally misleading information as misinformation requires additional consideration. Misleading information is different from information that is factually wrong. It can be created in pursuit of the truth and contain both facts and inaccuracies, without awareness of the latter. It is important to caution against considering or labeling this type of information misinformation. If "misinformation" does include information that is misleading, but which is created unintentionally, then this would pose problems for deciding what sources of information are authoritative and reliable versus those that are not, in addition to what sources should be trusted. Humans are fallible. They make mistakes, some of which can unintentionally create misleading information. Yet should such creations be labeled misinformation? Everything from academic works to news stories contains content that can be misleading even if it is not devoid of fact.

Four fitting examples speak to this point, both from traditional news sources and social media. First, in 2021 the BBC reported that Iran's Supreme Leader Ali Khamenei had apparently called for an attack on Donald Trump in revenge for the killing of Iran's military commander Qasem Soleimani.[12] Just one day after the BBC article's publication, *The Diplomat* issued an article about placing trust in false information, pointing out that this news story was based on a Twitter post from "a fake account which closely resembled Khamenei's."[13]

A second example is that when the polls closed in the 2017 UK general election, a statistic began to circulate that the turnout for voters aged 18-25 was 72% — except there "was just one problem: no one seemed to have the data to back any of this up. Not that this stopped news

visited Jan. 30, 2023); *Misinformation*, COLLINS ENG. DICTIONARY, https://www.collinsdictionary.com/us/dictionary/english/misinformation (last visited Jan. 30, 2023).

[11] *See, e.g.*, Marko Milanovic & Michael N. Schmitt, *Cyber Attacks and Cyber (Mis)information Operations During a Pandemic*, 11 J. NAT'L SEC. L & POL'Y 247, 266 (2020).

[12] *Iran's Supreme Leader Makes Online Threats to Attack Golfing Trump*, BBC (Jan. 22, 2021), https://www.bbc.com/news/world-middle-east-55765516.

[13] Abhijan Rej, *Faith in Fakes: What a "Khamenei Tweet" About Trump Actually Tells Us*, DIPLOMAT (Jan. 23, 2021), https://thediplomat.com/2021/01/faith-in-fakes-what-a-khamenei-tweet-about-trump-actually-tells-us/.

outlets from repeating the claims, all citing either unverified tweets or each other as sources."[14]

A third example is a *New York Times* article on the International Criminal Court and its jurisdiction over the situation in Palestine.[15] According to Kevin Jon Heller, the article made claims that were "simply false," and "[f]alse-false, not debatably false."[16]

Finally, an unsupported zombie statistic that has been pervasive for decades stems from a United Nations development report, stating that 70% of people living in poverty are women.[17] Duncan Green branded the statistic "dodgy" and Caren Gown labeled it "false" because there is a lack of sex-disaggregated data on this issue.[18] While apathy and/or trust are part of deciding if these mistakes were unintentional, they provide a snippet of the problems that come with categorizing misleading information as misinformation.

These four examples also call attention to what human beings know as opposed to what they think they know. David Hume argued that no human could be certain about anything when drawing inferences from observable experiences, as the "problem of induction" casts certainty as an "impossible ideal."[19] Why? Because things change over time and space. Marko Milanovic has noted that when facing the problem of misinformation,

> [W]e need to appreciate fully the contingent and mediated nature of the information that we believe to be accurate. We all need to put our trust in some authoritative sources of knowledge that we do not ourselves directly possess. I do not really know, through direct observation, that hundreds of people died in Iran due to misinformation-induced alcohol poisoning. I know this by choosing to believe reports from various media organizations that I consider to be reliable.[20]

---

[14] CAROLINE C. PEREZ, INVISIBLE WOMEN: DATA BIAS IN A WORLD DESIGNED FOR MEN 224 (2019).

[15] Isabel Kershner, *I.C.C. Rules It Has Jurisdiction to Examine Possible Israel War Crimes*, N.Y. TIMES (Feb. 5, 2021), https://www.nytimes.com/2021/02/05/world/middleeast/icc-israel-war-crimes.html.

[16] Kevin J. Heller, *The Gray Lady Botches Judge Kovács' Dissent*, OPINIO JURIS (Feb. 10, 2021), http://opiniojuris.org/2021/02/10/the-gray-lady-botches-judge-kovacs-dissent/.

[17] U.N. DEV. PROGRAMME, HUMAN DEVELOPMENT REPORT, at iii, 4, 36 (1995).

[18] PEREZ, *supra* note 14, at 224-25.

[19] STEPHEN BUCKLE, HUME'S ENLIGHTENMENT TRACT: THE UNITY AND PURPOSE OF AN ENQUIRY CONCERNING HUMAN UNDERSTANDING 151-69 (2004).

[20] Marko Milanovic, *Viral Misinformation and the Freedom of Expression: Part III*, EJIL TALK (Apr. 14, 2020), https://www.ejiltalk.org/viral-misinformation-and-the-freedom-of-expression-part-iii/.

Trust in sources of information is at least equally significant as individual knowledge to the spread of misinformation. Kate O'Regan has suggested that "the objective truth of a proposition may well not be the most significant factor in determining whether it comes to be accepted as true."[21] The blend of levels of trust that a person holds in sources of information, and their knowledge in relation to those sources, appears to impact whether that person will ultimately accept accurate information more than inaccurate information.[22]

Another factor to consider in attempting to understand the contours of misinformation is "bullshit"—information that is neither fact nor fiction, but somewhere in-between, or perhaps even beyond. Calling out bullshit is important to counteract the practice of propagating opinions that attempt to depict reality based on personal preferences, veil the truth, and make it harder to identify what needs addressing, why, and how. Lying—providing intentionally false information—implicitly acknowledges that the truth matters in some way. Otherwise, why lie? Bullshit, however, according to Harry Frankfurt, is a greater enemy of the truth than lies.[23] It can be hard to recognize and, consequently, to address, thus contributing to the spread of misinformation because the ability to misrepresent the truth is part of the reason why opinions can be passed off as facts.[24]

The mixed understandings about misinformation show that its occurrence can come about as a result of disinformation. What begins as the creation of information that is knowingly false, which is then intentionally shared with that knowledge, can result in that same source being shared further by those without the knowledge that it contains content contrary to fact. Disinformation may therefore become misinformation over time. For example, one actor may create and share false electoral results with the intent to disrupt a democratic process, but another actor may share that source without such intent but instead to express concern over a matter about which they are confused. Similarly, misinformation can turn into disinformation depending on the intent of the actor who shares it. The author of such information may not have been aware that it contained falsehoods, or indeed another actor that shared it. Yet a further actor upon becoming aware of the falsehoods contained

[21] Catherine O'Regan, *Hate Speech Online: an (Intractable) Contemporary Challenge?*, 71 CURRENT LEGAL PROBLEMS 403, 412 (2018).

[22] Stephen Lewandowsky, et al., *Misinformation and Its Correction: Continued Influence and Successful Debiasing,* 13 PSYCHOLOGICAL SCI. IN THE PUBLIC INTEREST 106 (2012).

[23] HARRY G. FRANKFURT, ON BULLSHIT 18 (2005).

[24] James B. Schreiber, *Be Careful! That is Probably Bullshit! Review of Calling Bullshit: The Art of Skepticism in a Data-Driven World by Carl T. Bergstrom and Jevin D. West,* 14 NUMERACY 2 (2021).

within that source may share it with an intent to deceive others. This blurs the line between whether that source is now misinformation or disinformation.

The other component of this changeability is the lack of knowledge regarding whether something is factual. This sets individual ignorance apart from the occurrence of disinformation, as intent implies knowledge of whether information is believed to be true. But being uninformed is not the same as being misinformed, even though being uninformed can contribute to the spread of misinformation. In the words of Denzel Washington, "If you don't read the newspaper, you're uninformed. If you do read it, you're misinformed."[25]

The boundaries of what constitutes misinformation may also overlap with those constituting disinformation. A distinguishing feature between the two is intent, or perhaps more accurately, the need or lack thereof to determine intent. Although establishing intent can be difficult to the extent that it is often inferred from the overall circumstances of a case, where an actor making a judgment on a particular set of facts can impute intent via inference even if it is not explicit,[26] the process of classifying misinformation does not need to undertake any such assessment of intent.

Misinformation is therefore at least information that is false as a matter of fact and that is created or shared without knowing that it is false or contains falsehoods. Yet whether it also includes unintentionally misleading information is difficult to say, and it may not be possible to settle what is acceptable in this respect. Lawyers, for example, are trained to use language in a way that conveys information that may not technically be wrong, but can lead to assumptions that it is right. Comedians supplement truths with satire, meaning that figurative or ironic expressions can be interpreted literally.[27] Media outlets update information in their coverage of unfolding events, meaning information that may have been accurate at one point in time may no longer be, such as death tolls from disasters and case numbers of a disease.

As such, this article is less concerned with misleading information or the sources of it than it is with false information and its unintentional

[25] Elyse Samuels, *Denzel Washington Calls upon Journalists to Tell the Truth*, THE WASH. POST (Dec. 14, 2016), https://www.washingtonpost.com/videoentertainment/denzel-washington-calls-upon-journalists-to-tell-the-truth/2016/12/14/b218db8e-c248-11e6-92e8-c07f4f671da4_video.html.

[26] ROBIN A. DUFF, INTENTION, AGENCY AND CRIMINAL LIABILITY: PHILOSOPHY OF ACTION AND THE CRIMINAL LAW 28 (1990).

[27] Dannagal Young, *Can Satire and Irony Constitute Misinformation?*, MISINFORMATION AND MASS AUDIENCES 124 (B. G. Southwell, E. A. Thorson, & L. Sheble ed., 2018).

spread by people that encounter it. This misinformation has become an associative characteristic of social media, one currently undergoing treatment.

B. The Particular Problem of Social Media

Misinformation on social media is widespread and predicted to worsen.[28] Three distinctive features of social media that set it apart from other media through which people receive and impart information are its scale, speed, and bespoke tailoring of information to precise individuals.[29] Social media platforms reach an estimated one-in-three people throughout the world.[30] Although the companies supplying these platforms apparently struggle to count the number of users they actually have, it is reported that at least two billion people are active users of Facebook alone.[31] The reasons behind why particular people use particular platforms are varied, but the streams of information flowing through these platforms are constant.[32] This continuous river of content containing facts, fiction, and everything else is instantaneous, both in terms of access and contribution.

Users on a particular platform can both be subjected to content the company operating that platform knows is more likely to increase their engagement, and subject other users to content they create—although the platform ultimately determines the visibility and reach of user-generated content.[33] Whether they come to share their own sources or to subject themselves to sources generated by other users and the hosting platform, people that are used by social media companies to gather monetizable data can be fed misinformation every time they hook up to their favored platforms for a hit.

---

[28] Anjana Susarla, et al., *What Will 2022 Bring in the Way of Misinformation on Social Media? 3 Experts Weigh in*, THE CONVERSATION (27 Dec. 27, 2021), https://theconversation.com/what-will-2022-bring-in-the-way-of-misinformation-on-social-media-3-experts-weigh-in-173952; Janna Anderson & Lee Rainie, *The Future of Truth and Misinformation Online*, PEW RSCH. CTR. (19 Oct. 19, 2017), https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/.

[29] *See* James Grimmelmann, *The Platform is the Message*, 2 GEO. L. TECH. REV. 217, 224-230 (2018).

[30] Estaban Ortiz-Ospina, *Are Facebook and Other Social Media Platforms Bad for Our Well-Being?*, OUR WORLD IN DATA (Sept. 9, 2019), https://ourworldindata.org/social-media-wellbeing.

[31] Hannah Murphy, *Facebook Confronts Growth Problems as Number of Young Users in US Declines*, FIN. TIMES (Oct. 22, 2021), https://www.ft.com/content/4304f14a-1b06-46d8-a066-42bb1b3c200c.

[32] Tara C. Marshall, et al., *Intellectual, Narcissistic, or Machiavellian? How Twitter Users Differ from Facebook-Only Users, Why They Use Twitter, and What They Tweet About*, 9 PSYCH. OF POPULAR MEDIA 14 (2020).

[33] *See* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 197-327 (2019).

In addition to the instantaneous nature of information on social media that can reach billions of people in a matter of minutes, the type of information that generates more engagement is key. In this respect, research shows that misinformation spreads faster and further on social media than information considered to be true.[34] There are multiple factors that lead to such occurrences, but a recurrent one is a source's ability to elicit an emotional response.[35] When one person knows another well, they can attempt to generate a particular emotional response that is likely to result in further interaction (winding up a friend to share a laugh is an example). The algorithms on social media platforms do something similar, with their ability to generate further interaction being as—if not more—successful. Their success derives from the amount of user data they hold, as well as the practice of selectively presenting sources to users that are predicted to engage with them, and through those users' (foreseeable) responses, even more engagement is generated from other users.[36]

However, a key difference between a close personal contact and these algorithms is that the latter appears to equate engagement with what a person values.[37] Building this conflation into algorithmic design is arguably flawed because the algorithms fail to recognize that high engagement levels with a particular piece of content does not necessarily mean that the users engaging with that content value it or even want to be made aware of it. Twitter conducted a study showing that its algorithms amplify particular types of content because the "ranking content on the Home timeline is influenced by the output of deep learning models, trained to predict various types of engagements with Tweets (likes, reTweets, replies, etc)."[38]

---

[34] Soroush Vosoughi, et al., *The Spread of True and False News Online,* 359 SCI. 1146 (2018); Nir Grinberg, et al., *Fake News on Twitter during the 2016 U.S. Presidential Election*, 363 SCI. 374 (2019).

[35] Jonah Berger, *Arousal Increases Social Transmission of Information*, 22 PSYCH. SCI. 891 (2011); Kim Peters, et al., *Talking about Others: Emotionality and the Dissemination of Social Information*, 39 EUR. J. OF SOC. PSYCH. 207 (2009); Ellen M. Cotter, *Influence of Emotional Content and Perceived Relevance on Spread of Urban Legends: A Pilot Study*, 102 PSYCH. REP. 623 (2008); Chip Heath, et al., *Emotional Selection in Memes: The Case of Urban Legends*, 81 J. OF PERSONALITY AND SOCIAL PSYCH. 1028 (2001).

[36] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 179-185 (New York: Crown Publishers 1st ed., 2016).

[37] *See* Gilad Edelman, *Facebook Quietly Makes a Big Admission*, WIRED (Aug. 31, 2021), https://www.wired.com/story/facebook-quietly-makes-big-admission-political-content/.

[38] Ferenc Huszár, et al., *Algorithmic Amplification of Politics on Twitter*, TWITTER BLOG, SI p. 3 (Oct. 21, 2021) https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2021/rml/Algorithmic-Amplification-of-Politics-on-Twitter.pdf; *See also*__Ferenc Huszár, et al., *Algorithmic Amplification of Politics on Twitter,* 119 PROCEEDINGS OF THE NAT'L. ACADEMY OF SCI., Dec. 21. 2021, https://www.pnas.org/doi/epdf/10.1073/pnas.2025334119.

According to Twitter, their algorithms assess the following criteria when predicting what would be engaging to each user:

> - The Tweet itself: its recency, presence of media cards (image or video), total interactions (e.g., number of Retweets or likes);
>
> - The Tweet's author: your past interactions with this author, the strength of your connection to them, the origin of your relationship;
>
> - You: Tweets you found engaging in the past, how often and how heavily you use Twitter.[39]

While these algorithms can be updated to better reflect users' basic values, or may change in other ways due to Twitter's recent change in ownership,[40] the market-driven ethos underlying their creation incentivizes conduct aimed at maximizing utility (which for social media companies currently appears to be user-engagement) without much regard for the substance behind each form of such engagement and its resulting costs, except perhaps those incurred by the individual companies to their brand and bottom line. As platforms' algorithms are designed so that each individual user has sources tailored to them to increase engagement, the content of those sources does not matter so much *per se* in serving this goal, but in whether it is likely to generate an emotional reaction. Misinformation that is evocative or provocative for large quantities of users will likely be spread far and wide without users knowing they are interacting with false information, even if such misinformation originates in a small number of sources and then attains outsized reach.

One of the challenges for governance of social media is combating the spread of misinformation in a way that accounts for the possibility that the algorithmic design practices of these platforms, which are aimed at increasing user-engagement, are unlikely to change much. This can be said with some confidence so long as platform design remains unregulated, continues to be a means of extracting more user data, and can generate more revenue through the exploitation of that data.[41]

Some have called for outright bans on practices such as targeted advertising and other measures that would render social media platforms

---

[39] Nicolas Koumchatzky & Anton Andryeyev, *Using Deep Learning at Scale in Twitter's Timelines*, TWITTER BLOG (May 9, 2017), https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.

[40] Ivan Mehta, *A List of Features Elon Musk Has Promised to Bring to Twitter*, TECH CRUNCH (Nov. 8, 2022), https://techcrunch.com/2022/11/08/a-list-of-features-elon-musk-has-promised-to-bring-to-twitter/.

[41] *See generally* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2019).

liable for harmful outcomes that stem from user-generated content.[42] The regulation of social media companies, particularly with respect to what information is shared on their platforms, is a challenge that new policies are tasked with addressing. Yet their implementation depends on overcoming the lobbying efforts of social media companies or appealing to the companies themselves to improve their self-regulatory efforts.[43]

Although the calls to regulate social media platforms can be forceful, they outnumber ideas and agreement about how precisely to go about it. What specific measures should be taken to stem the flow of misinformation on social media platforms? One possibility is targeted user-facing behavioral interventions on newsfeeds, namely operationalizations of "nudging," a concept that has imbued thinking across numerous fields.

II. NUDGING ON NEWSFEEDS

The idea of nudging has existed since at least the 1940s and was previously referred to as "behavioural engineering."[44] The research surrounding this idea is rich and its antecedents are extensive.[45] What can these insights offer initiatives aimed at managing misinformation on social media platforms? Answers to this question become apparent when examining the workings of social media newsfeeds and how they might be adapted in light of related research across behavioral economics, cognitive psychology, and human rights.

A. The Premise Behind Nudging

The purpose here is not to propose a comprehensive theory of nudging, which, though called for in the literature to help settle a number of

[42] Clothilde Goujard, *CEOs Make Final Push to Ban Targeted Ads*, POLITICO (Jan. 13, 2022), https://www.politico.eu/article/activist-ceo-mep-crack-down-targeted-ads-vote-digital-services-act-2/.

[43] Dipayan Ghosh, *Are We Entering a New Era of Social Media Regulation?*, HARV. BUS. REV. (Jan. 14, 2021); Louise Matsakis, *Facebook's Targeted Ads Are More Complex Than It Lets On*, WIRED (Apr. 25, 2018), https://www.wired.com/story/facebooks-targeted-ads-are-more-complex-than-it-lets-on/.

[44] Magda Osman, *Nudges: Four Reasons to Doubt Popular Technique to Shape People's Behaviour*, THE CONVERSATION (Jan. 10, 2022), https://theconversation.com/nudges-four-reasons-to-doubt-popular-technique-to-shape-peoples-behaviour-174359.,

[45] *See, e.g.*, Cass R. Sunstein, et al., *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471 (1998); Justin Fox, *From "Economic Man" to Behavioral Economics*, HARV. BUS. REV. 75, 78-85 (May 2015); RICHARD H. THALER, MISBEHAVING: THE MAKING OF BEHAVIORAL ECONOMICS (2016).

debates, goes beyond the aims of this article.[46] However, the general premise of nudging requires clarification. Doron Teichman and Eyal Zamir describe nudges as "regulatory tools that use psychological insights to design the decision-making environment in a way that promotes certain choices."[47] The ways human behavior regarding choice can be influenced by a nudge include correcting misapprehensions through additional information, changing how choices are presented to people, and implementing default options.[48] Nudges are based on an underlying assumption that they will alter human behavior confined to choice without excluding the opportunity to choose between various options.[49] This approach to decision-making is one that promises to preserve "freedom of choice but that authorizes both private and public institutions to steer people in directions that will promote their welfare."[50]

The framework informing much of this discourse was popularized by the work of Daniel Kahneman.[51] His work's central tenet is that the processes of human reasoning, decision-making, and judgment are determined by two cognitive mechanisms: System 1 and System 2 thinking.[52] System 1 consists of quick thinking that is automatic, interactional, and guided by heuristics, whereas System 2 consists of comparatively slower thinking that is more conscious, analytical, and guided by weighted reasoning.[53] While the efficacy of nudges in general appears to be somewhat fleeting, their use is intended to interface with

[46] *See generally* Till Grüne-Yanoff & Ralph Hertwig, *Nudge Versus Boost: How Coherent are Policy and Theory?*, 26 MINDS AND MACHINES 149 (2016); David Trafimow, *The Role of Auxiliary Assumptions for the Validity of Manipulations and Measures*, 22(4) THEORY & PSYCHOLOGY 486 (2012); Luc Bovens, *The Ethics of Nudge*, in PREFERENCE CHANGE: APPROACHES FROM PHIL., ECON. AND PSYCHOL. 207 (Tille Grüne-Yanoff and Sven Ove Hansson eds., 2009).

[47] Doron Teichman & Eyal Zamir, *Nudge Goes International*, 30 EUR. J. OF INT'L. L. 1263, 1266 (2019).

[48] Yiling Lin, et al., *Nudge: Concept, Effectiveness, and Ethics*, 39 BASIC AND APPLIED SOC. PSYCHOL. 293, 293 (2017).

[49] RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 6 (Penguin Books, 2009).

[50] Richard H. Thaler & Cass R. Sunstein, *Libertarian Paternalism*, 93 THE AM. ECON. REV. 175, 179 (2003).

[51] DANIEL KAHNEMAN, THINKING, FAST AND SLOW (Farrar, Straus and Giroux, 2011); *see also* Andreas T. Schmidt & Bart Engelen, *The Ethics of Nudging: An Overview*, 15 PHIL. COMPASS 1 (2020); Jessica L. Roberts, *Nudge-Proof: Distributive Justice and the Ethics of Nudging*, 116 MICH. L. REV. 1045 (2018); CASS R. SUNSTEIN, THE ETHICS OF INFLUENCE: GOVERNMENT IN THE AGE OF BEHAVIORAL SCIENCE (2016).

[52] Kahneman, *supra* note 52; *see also* THOMAS GILOVICH, DALE GRIFFIN & DANIEL KAHNEMAN, HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT 49 (2002).

[53] KEITH E. STANOVICH, WHO IS RATIONAL? STUDIES OF INDIVIDUAL DIFFERENCES IN REASONING (1999).

System 1.[54] Socioeconomic and sociopsychological factors influence how such engagement plays out. Depending on the method of measurement, these factors can result in a variety of outcomes, even if some studies find that nudges are an effective behavioral change tool to facilitate "personally and socially desirable choices."[55]

Nudging does not always work[56] and can have harmful distributional effects.[57] The "efficacy of interventions may vary across contexts: what works well in one situation or with one group of people may be of limited use in different settings or with different cultural groups."[58]

The process of creating choice architecture involves "(re)designing the physical, social, or psychological environment in which people make decisions" for the purpose of guiding decision-making and influencing conduct of those people toward a particular result,[59] the predictability of which can change depending on the design, the domain, and the person(s) at hand. Given the heterogeneity in research on nudging and its results,[60] examining digital nudging on social media newsfeeds provides the benefit of working toward clarity regarding a matter that is prominent on the agendas of policymakers and industry leaders tasked with tackling the use and influence of algorithms and the spread of misinformation.

B. Digital Nudging on Social Media: A User-Choice Intervention?

At least two types of digital nudges can be implemented on social media newsfeeds: (1) fact-check alerts and labeling; and (2) alternative source presentation. Both have the potential to reduce the spread of misinformation, whether they are employed individually or in combination.

---

[54] Tina A. G. Venema, et al., *When in Doubt, Follow the Crowd? Responsiveness to Social Proof Nudges in the Absence of Clear Preferences*, 11 FRONTIERS IN PSYCHOL. 1, 11 (2020); Dennis Hummel & Alexander Maedche, *How Effective is Nudging? A Quantitative Review on the Effect Sizes and Limits of Empirical Nudging Studies*, 80 J. OF BEHAV. AND EXPERIMENTAL ECON. 47, 47 (2019); Björn Meder, Nadine Fleischhut & Magda Osman, *Beyond the confines of choice architecture: A critical analysis*, 68 J. OF ECON. PSYCHIOL. 36, 42 (2018).

[55] Stephanie Mertens, et al., *The Effectiveness of Nudging: A Meta-Analysis of Choice Architecture Interventions across Behavioral Domains*, 119 PROC. OF THE NAT'L ACAD. OF SCI. 1, 8 (2020).

[56] Cass R. Sunstein, *Nudges that Fail* 1 BEHAV. PUB. POL'Y 1, 4 (2017).

[57] Cass R. Sunstein, *The Distributional Effects of Nudges*, 6 NAT. HUM. BEHAV. 9, 9 (2022).

[58] Nichola J. Raihani, *Nudge Politics: Efficacy and Ethics*, 4 FRONTIERS PSYCHOL. 1, 1 (2013).

[59] Stephanie Mertens, et al., *The Effectiveness of Nudging: A Meta-Analysis of Choice Architecture Interventions Across Behavioral Domains*, 119 PROC. NAT'L. ACAD. SCI. 1, 1 (2022); *See also* RICHARD H. THALER, ET AL., *Choice Architecture*, *in* THE BEHAV. FOUNDATIONS OF PUB. POL'Y 428 (Eldar Shafir ed., 2012).

[60] Cristina Mele, et al., *Smart Nudging: How Cognitive Technologies Enable Choice Architectures for Value Co-Creation,* 129 J. BUS. RSCH. 949 (2021).

These are distinguishable from nudges that are implemented with the aim of facilitating the spread of information, such as promoting "police or government accounts so that accurate information is disseminated as quickly as possible."[61] The two digital nudges examined below are aimed at *decreasing* the spread of misinformation, not content promotion or increasing the spread of (hopefully accurate) information in response to situations such as disasters and emergencies.[62] After initially exploring these two digital nudges, this section will then compare them to other measures that aim to address misinformation on social media before discussing the risks and shortcomings associated with their use.

1. Fact-Check Alerts and Labeling

Fact-check alerts are digital nudges that do what they say on the tin. When a source of information appears on a particular social media platform, it can be labeled in a variety of ways to indicate that its contents are suspect. These notifications can take the form of information panels, pop-ups, and tags, which are becoming ingrained on social media platforms.[63] Right off the bat there are two significant features of this practice that concern the data on which digital nudges rely when they are created. The first is who the arbiters of factual truths are and who decides that they should have this role (and what it will entail).[64] At the moment, social media companies make these decisions. The second is fact-checkers' respective agendas, if any.[65] While transparency determines the extent to which any such agendas can be known, once again, trust is key. However, while noteworthy, these two features of fact-checking are not the focus here. Instead, the focus is on how this practice works in the form of a digital nudge and its resulting effect on the spread of misinformation.

---

[61] Chris Meserole, *How Misinformation Spreads on Social Media—And What To Do About It*, LAWFARE (May 9, 2018), https://www.lawfareblog.com/how-misinformation-spreads-on-social-media-and-what-to-do-about-it.

[62] *See* Milad Mirbabaie, et al., *Digital Nudging in Social Media Disaster Communication*, 23 INFO. SYS. FRONTIERS 1097 (2021).

[63] *Meta's Third-Party Fact-Checking Program*, FACEBOOK: META FOR MEDIA, https://www.facebook.com/journalismproject/programs/third-party-fact-checking (last visited Feb. 8, 2023); Keith Coleman, *Introducing Birdwatch, a Community-Based Approach to Misinformation*, TWITTER: BLOG (Jan. 25, 2021), https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[64] Hunt Allcott & Matthew Gentzkow, *Social Media and Fake News in the 2016 Election*, 31 J. ECON. PERSP. 211, 233 (2017).

[65] Maria Haigh, et al., *Stopping Fake News: The Work Practices of Peer-to-Peer Counter Propaganda*, 19 JOURNALISM STUD. 2062 (2017).

A hypothesis empirically tested by Elmie Nekmat is that exposure to fact-check alerts on newsfeeds warning users of misinformation will lower the likelihood of sharing compared to non-exposure.[66] The methodology behind this study clarifies that participants "were randomly exposed to a fact-check alert that popped up next to the news report to warn of the informational inaccuracy contained in the news report."[67] The results show that participants "were less likely to share the news when exposed to the fact-check alert . . . compared to non-exposure," thus supporting the hypothesis.[68]

There are two key components that explain the inverse correlation between using this digital nudge and the likelihood of misinformation being shared. These are: (1) how people perceive information presented to them; and (2) how they respond to that perception in a particular context. Providing a label of some sort to raise the users' awareness about a source of information is believed to influence how people perceive it. By placing a label accompanying content that is posted on a newsfeed, flagging it in the way of a warning, it is assumed that the user's perception of that content will change compared to if no label accompanied it. Yet whether or not this change of perception occurs, it is how a user responds to the label that ultimately determines whether misinformation will spread any further than the user exposed to the fact-check alert. By tapping into the biases and heuristics that saturate System 1 thinking, research suggests that users warned about misinformation will be dissuaded from sharing it.[69] The rationale here is that an alert attached to a particular source perceived by a user when scrolling their newsfeed will highlight a potential loss for that user if they were to share the source—that loss being damage to their personal reputation. In order to maintain their perceived status among connections on a particular social media platform, the user is thus nudged toward not sharing the potentially false source.

This procedure of fact-checking and labeling might be influenced by how frequently users interact with their social media connections. The number of interactions people have with their connections is linked to how much their behavior changes in a given context.[70] Increases in the type of

---

[66] Elmie Nekmat, *Nudge Effect of Fact-Check Alerts: Source Influence and Media Skepticism on Sharing of News Misinformation in Social Media*, 6 SOC. MEDIA SOC'Y 1, 3 (2020).

[67] *Id*. at 5.

[68] *Id*. at 7.

[69] *See* Daniel Kahneman, et al., *Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias*, 5 J. ECON. PERSP. 193 (1991).

[70] ALEX PENTLAND, SOCIAL PHYSICS: HOW SOCIAL NETWORKS CAN MAKE US SMARTER 69 (2015).

interactions among connections where fact-check alerts are used could therefore counter the problem of misinformation spreading across users that would ordinarily increase through such interactions when fact-check alerts are not used.

Recall social media platforms' aim of increasing user engagement. Digital nudging flips this problem on its head. Alex Pentland found that since "exchanges between people are of enormous value to the participants, we can leverage those exchanges to generate social pressure for change. Engagement—repeated cooperative interactions among members of the community—brings movement toward cooperative behavior."[71] Social network dynamics can become a medicine for the market-driven incentives influencing social media companies, with the potential to reduce the resulting net negative effects of misinformation through the very goal of user-engagement that companies seek to maximize. This practice may be of more use when connections on social media see each other in person as well, meaning the combined digitized and in-person interactions are likely significant for the functionality of digital nudging consisting of fact-check alerts.[72]

A crucial difficulty is users having connections that hold the same or similar views, and groups where echo chambers result in misinformation being met with collective approval.[73] This does not mean that individual users cannot broaden or burst their own informational bubbles without interference from another party.[74] What it does mean is users become overconfident in their position on a subject, rooted to it, even in the face of its fallacies.[75] Confirmation bias helps explain why everyone from investment bankers to political pundits get things wrong, especially when making predictions.[76] In the words of Marina Hyde, "It's actually quite difficult to find people who are more wrong on a regular basis than

---

[71] *Id*.

[72] *Id*. at 70-75.

[73] Ray Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 REV. GEN. PSYCHOL. 175 (1998); Ray Nickerson, *Confirmation Bias: A Psychological Phenomenon that Helps Explain Why Pundits Got It Wrong*, THE CONVERSATION (Nov. 22, 2016), https://theconversation.com/confirmation-bias-a-psychological-phenomenon-that-helps-explain-why-pundits-got-it-wrong-68781.

[74] Christopher Seneca, *How to Break Out of Your Social Media Echo Chamber*, WIRED (Sept. 17, 2020), https://www.wired.com/story/facebook-twitter-echo-chamber-confirmation-bias/.

[75] *See generally* MADSEN PIRIE, HOW TO WIN EVERY ARGUMENT: THE USE AND ABUSE OF LOGIC (Bloomsbury 2d ed. 2015) (2006).

[76] *See* DANIEL KAHNEMAN, THINKING, FAST AND SLOW 209-221 (2011); DANIEL KAHNEMAN, ET AL., NOISE: A FLAW IN HUMAN JUDGMENT 140-142 (2021); *See also generally* PHILIP E. TETLOCK, EXPERT POLITICAL JUDGMENT: HOW GOOD IS IT? HOW CAN WE KNOW? (2017).

newspaper columnists—but it's possible that economists do edge it."[77] While there may be other strong contenders for this title, Daniel Kahneman, Olivier Sibony, and Cass Sunstein point out a potential common denominator that it is those "blessed with clear theories about how the world works" who tend to be "the most confident and the least accurate."[78]

This combination of overconfidence, confirmation bias, and echo chambers helps explain why a digital nudge in the form of a fact-check alert may, at best, be minimally effective at reducing the spread of misinformation. People can be overconfident in their own ideas, which can be reinforced by others within their circle of connections. This emphasizes the importance for those that are used by social media platforms to follow accounts and share connections with those that have different opinions and interests from their own.[79] Attempting to replicate online the social aspects of human interaction and gravitation toward comfort that happens offline is a pernicious feature of social media. Nevertheless, even in the face of confirmation bias, Facebook claims that fact-check alerts work.[80] Regrettably, this claim does not speak to whether labeled content is shared. Users can and do share sources without ever actually reading them (beyond the headline that is visible on platforms without having to click through onto the linked content). Twitter has also stated that "prompts helped decrease Quote Tweets of misleading information by 29% so we're expanding them to show when you tap to like a labeled Tweet."[81] These claims are worth reflecting on in light of research providing "support for prior studies finding a negative effect of general warnings on belief in

[77] Marina Hyde, *Who Better Than Liz Truss to Lead a Country Whose Own Sewage Laps at Its Shores*?, THE GUARDIAN (Aug. 19, 2022), https://www.theguardian.com/commentisfree/2022/aug/19/liz-truss-lead-uk-sewage-leadership-marina-hyde.

[78] Kahneman, et al., *supra* note 77, at 141.

[79] *See also* Alexander Bor & Michael Bang Petersen, *The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis*, 116 AM. POL. SCI. REV. 1 (2022) (discussing how the dynamic can change between users should hostility form part of the misinformation sharing process).

[80] Mark Zuckerberg, FACEBOOK (Apr. 29, 2020, 4:32 PM), https://www.facebook.com/zuck/posts/10111862059015441 ("For other types of misinformation, we partner with independent fact-checkers, who have marked more than 4,000 pieces of content related to Covid as false, which has resulted in more than 40 million warning labels being seen across our services. And we know these work because 95% of the time when someone sees a warning label, they don't click through to view that content").

[81] Twitter Support (@TwitterSupport), TWITTER (Nov. 23, 2020, 5:33 PM), https://twitter.com/TwitterSupport/status/1331018136712261632.

misinformation," while showing "tags modestly reduce belief in false news."[82]

How a label is presented appears to be decisive in whether a fact-check digital nudge works. The apparent ineffectiveness of some fact-check digital nudges may stem from their design and placement on a newsfeed relative to the content in question. A wee exclamation point within a colorful triangle next to a statement such as "Misleading" may go unnoticed by users.[83] An "important development has been a new format of warning that interrupts users' actions and forces them to make a choice about whether to continue."[84] These are called interstitial webpages. Having such measures on social media newsfeeds adds "friction" into user interfaces, meaning users are prompted to consider "whether they really want to post certain content," which may discourage sharing sources of misinformation.[85] This alteration may prove to be more effective than other versions of fact-check alerts. An overarching takeaway is that social media platforms are using these digital nudges already, appear to be developing them further, and may implement them more widely in the future, particularly in advance of or during certain events, such as elections and armed conflicts, respectively, where there are spikes in the amount of misinformation present online.[86]

As spotting misinformation on social media can be challenging, initiatives attempting to address this problem through fact-check-based digital nudging have the potential to make this process easier. But what else in the collection of digital nudges can help manage misinformation on social media? Can informational bubbles riddled with confirmation bias be altered, if not burst?

---

[82] Katherine Clayton, et al. *Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media*, 42 POL. BEHAVIOR 1073, 1090-1091 (2020).

[83] *See* Jane Manchun Wong (@wongmjane) TWITTER (May 31, 2021, 5:26 AM), https://twitter.com/wongmjane/status/1399311420794105862.

[84] Ben Kaiser, et al., *Warnings That Work: Combating Misinformation Without Deplatforming*, LAWFARE (Jul. 23, 2021), https://www.lawfareblog.com/warnings-work-combating-misinformation-without-deplatforming.

[85] Molly K. Land & Rebecca. J. Hamilton, *Beyond Takedown: Expanding the Toolkit for Responding to Online Hate*, *in* PROPAGANDA AND INTERNATIONAL CRIMINAL LAW: FROM COGNITION TO CRIMINALITY 143, 152-153 (Predrag Dojčinovic ed., 2019).

[86] Chengcheng Shao, et al., *Anatomy of an Online Misinformation Network*, PLOS ONE 1 (Apr. 27, 2018).

2. Alternative Source Presentation: A Conduit for Informational Osmosis

The second digital nudge that can be used on social media newsfeeds presents alternative sources of information to those posted by a user or platform. Like fact-check alerts, alternative source digital nudges are self-descriptive. Their implementation consists of displaying alternative sources of information that relate to the posted content directly below or to the side of the original post,[87] or as an interstitial pop-up should a user click to read or share a source containing misinformation.

For example, say a user posts something linking to a source claiming that COVID-19 patients "should drink cow urine and chant Shiva mantras."[88] The possible alternative sources that could be presented alongside, underneath, or as an interstitial pop-up accompanying that post include: (1) a source discussing the lack of evidence regarding possible treatments for COVID-19; (2) another source linking to the World Health Organization page on how to report misinformation; and/or (3) a source from the ministry of health in the state where the user is physically located, containing information on what to do upon presenting with symptoms of COVID-19. The idea behind this digital nudge is that "a user is given the opportunity to forge their own opinion by reading from multiple sources."[89] As users can be presented with alternative framings of the same information, in addition to accurate sources that concern the content of original posts, this measure appears to be geared toward kick-starting System 2 thinking, overriding the default System 1, even if briefly.

As users are directed toward reading from more than one source on a subject, this digital nudge has the potential to counteract confirmation bias at scale. In addition to social media platforms producing large amounts of information at great speeds, providing "an extra edge over other sources of knowledge," there is again the problem of online networks that limit users' exposure to different viewpoints.[90] Using social media "allows us to construct and prune our social networks, to surround ourselves with others

---

[87] *See* Calum Thornhill, et al., *A Digital Nudge to Counter Confirmation Bias*, 2 FRONTIERS IN BIG DATA 1, 1-4 (Jun. 6, 2019).

[88] *Coronavirus: Can Cow Dung and Urine Help Cure the Novel Coronavirus*, TIMES OF INDIA (Feb. 5, 2020, 11:30 IST), https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/coronavirus-in-india-can-cow-dung-and-urine-help-cure-the-novel-coronavirus/articleshow/73952691.cms.

[89] Calum Thornhill, et al., *A Digital Nudge to Counter Confirmation Bias*, 2 FRONTIERS BIG DATA 1, 3 (2019).

[90] CAILIN O'CONNOR & JAMES OWEN WEATHERALL, THE MISINFORMATION AGE: HOW FALSE BELIEFS SPREAD 16 (2019).

who share our views and biases, and to refuse to interact with those who do not. This, in turn, filters the ways in which the world can push back, by limiting the facts to which we are exposed."[91] Pushing back against misinformation is possible, but when considering the hours users spend on social media on a daily basis, on top of those hours spent tending to personal and professional commitments, the windows of time each day in which any user of social media will be exposed to alternative sources of information are possibly reduced compared to if they do not use social media.[92]

This is where the presumption that "more information" is a solution to misinformation on social media falls short. Cailin O'Connor and James Owen Weatherall point out that while "it might seem that the solution is more information, this view is too limited. We have more information than ever before . . . it is the abundance of information, shared in novel social contexts, that underlies the problems we face."[93] Navigability of the online information environment is particularly difficult, not only because of informational volume, but also because of digital design that is effective at catching and holding attention.[94]

Relatedly, informational overload combined with users' cognitive biases call into question the marketplace of ideas approach favored in U.S. caselaw.[95] Information that holds a basis in fact will not always reach people who receive their information through social media platforms, in particular because of echo chambers across social networks and algorithmic design practices geared toward increasing user engagement.[96] The combination of these factors means accurate information, no matter the amount, will not be capable of reducing concentrations of misinformation within the informational bubble of a social media user, unless accurate information reaches its target, thereby permeating that bubble. Figure 1 provides a representation of this point (not to scale).

---

[91] *Id.*

[92] *Social Media Fact Sheet*, PEW RSCH. CTR. (Apr. 7, 2021), https://www.pewresearch.org/internet/fact-sheet/social-media/ (last visited Feb. 1, 2023).

[93] O'Connor & Weatherall, *supra* note 91, at 18.

[94] Ulrik Lyngs, et al., *I Just Want to Hack Myself to Not Get Distracted: Evaluating Design Interventions for Self-Control on Facebook* (Apr. 25–30, 2020) (proceedings of the 2020 CHI Conference on Human Factors in Computing Systems).

[95] *See* Derek E. Bambauer, *Shopping Badly: Cognitive Biases, Communications, and the Fallacy of the Marketplace of Ideas*, 77 U. COLO. L. REV. 649, 673–703 (2006).

[96] *See* Petter Tōrnberg, *How Digital Media Drive Affective Polarization through Partisan Sorting*, 119 PNAS 1, 1–11 (2022) (discussing some aspects of social media use may actually break echo chambers, particularly where users are brought to interact with others holding different viewpoints).

x (big bubble) = Potential information available to a user (offline and online).

y (small bubble) = Information perceived by a user when on a social media platform.

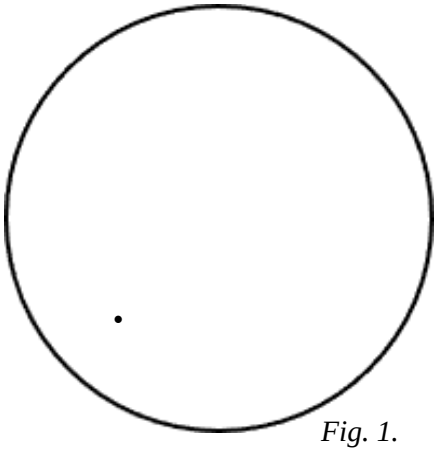

Fig. 1.

No matter how much of an increase occurs in the potential information available to a user (*x*), so long as it does not permeate the bubble of information perceived by that user when on a social media platform (*y*), more information cannot counteract misinformation within that informational bubble. In other words, *y* will remain unchanged by changes in *x* unless some method of informational osmosis is utilized whereby the contents of *x* pass through into *y*.

One delivery method that has the potential to provide for such a transfer is the alternative source nudge, as it functions as a conduit for information from *x* to permeate *y*. A key issue thus becomes the selection of sources to present as alternatives. It has been argued that "when people encounter a piece of information, they can check it against other knowledge to assess its compatibility."[97] This process is "effortful, and it requires motivation and cognitive resources," in addition to being influenced by a person's "affective response to new information," which supports "the assumption that information that is inconsistent with one's beliefs elicits negative feelings."[98] But what if an algorithm were designed to function on social media newsfeeds so that new information is automatically presented to a user when they encounter a source that contains misinformation, in a way attuned to their affectivity?

[97] Stephan Lewandowsky, et al., *Misinformation and its Correction: Continued Influence and Successful Debiasing*, 13 PSYCH. SCI. PUB. INT. 106, 112 (2012).

[98] *Id.*

A team of mathematicians recently created a statistical model that can detect misinformation with high levels of both accuracy and, arguably more importantly, explainability.[99] So, the first part of such an algorithmic design process is possible: being able to recognize misinformation and explain how that recognition occurred. As to the second part, what information should be presented to users as an alternative when they encounter a source of misinformation, the very practices of social media platforms provide an answer. Social media platforms know a lot about their users.[100] According to Shoshana Zuboff, people used by social media platforms are perhaps viewed as nothing more than raw material from which to extract data that concerns the innermost thoughts and feelings of users.[101] Whether it is Facebook, LinkedIn, Twitter, or another platform, data regarding their users is collected non-stop, all day, every day.[102] The datasets that social media platforms have on each of their users can be used to inform what alternative sources could be presented to users alongside original content.

Based on insights into these data mining practices, it can be reasoned that platforms have strong indications of what sources of information each user trusts. This means an algorithm can be designed that first detects sources of misinformation and, upon such recognition, presents alternative sources that do not contain falsehoods and which the specific user is likely to trust. This digital nudge would likely have more potential at countering confirmation bias than the use of fact-check alerts alone, as new information would be presented to users in a way that factors in their likely affectivity toward it, thereby guiding them in overcoming misapprehensions. As Briony Swire and Ullrich Ecker note, "[o]ne of the most effective methods of correcting misinformation is to provide an alternative factual cause or explanation to facilitate switching out the inaccurate information in an individual's initial situation model."[103] But the use of such alternative information needs to account for how it makes people feel in addition to what they think.

---

[99] CAITLIN MORONEY, ET AL., THE CASE FOR LATENT VARIABLE VS DEEP LEARNING METHODS IN MISINFORMATION DETECTION: AN APPLICATION TO COVID-19 422 (2021).

[100] ALEX PENTLAND, SOCIAL PHYSICS (2015); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016); SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2019).

[101] *See generally* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2019).

[102] Abby McCourt, *Social Media Mining: The Effects of Big Data In the Age of Social Media*, YALE L. SCH. (Apr. 3, 2018), https://law.yale.edu/mfia/case-disclosed/social-media-mining-effects-big-data-age-social-media.

[103] BRIONY SWIRE & ULLRICH ECKER, *Misinformation and Its Correction: Cognitive Mechanisms and Recommendations for Mass Communication*, *in* MISINFORMATION AND MASS AUDIENCES 195, 198 (2018).

The other factor to consider with this digital nudge compared to fact-checking and labeling is that social media companies do not become the arbiters of truth in a binary fashion. This is because there would be no explicit notification presented to a user that a source may be misinformation. Instead, different sources are presented to users that are not misinformation. The alternative source nudge is distinct in that it does not require making an explicit statement about what is true and what is false. In using it, social media companies themselves would therefore not be in the controversial position of deciding what is true or not, but instead would be encouraging wider reading on a subject by presenting the option to consider alternative viewpoints. This measure could therefore be part of social media platforms contributing to more diverse source consumption than at present.[104]

That said, as this digital nudge would need to be created in the form of an algorithm, humans creating the datasets informing its decision-making would need to initially categorize sources of information in terms of their relationship to truth (for example, true, false, misleading, unclear, or disputed). Automated processes function mostly as well as the information they are fed by humans, meaning biased, "noisy" human thinking introduces data that results in a biased algorithm with potentially low levels of explainability,[105] even if it cannot be a noisy decision-maker.[106] Those designing this digital nudge would need to exercise more care than such teams have done previously.[107]

Another salient feature of this digital nudge is that it has the potential to help address the floods of misinformation that appear on social media which lead to platforms implementing censorship measures.[108] This means when "reverse censorship" occurs, where large amounts of information are introduced in an attempt to counter measures that some may consider

---

[104] Amy Ross Arguedas, et al., *Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review*, REUTERS INST. FOR THE STUDY OF JOURNALISM (Jan. 19, 2022), https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review; The Reuters Institute for the Study of Journalism at the University of Oxford is funded by the Google News Initiative and the Facebook Journalism Project, among other funders. *See Our Funders*, REUTERS INSTITUTE FOR THE STUDY OF JOURNALISM, https://reutersinstitute.politics.ox.ac.uk/our-funders.

[105] *See* Aislinn Kelly-Lyth, *Challenging Biased Hiring Algorithms*, 41 OXFORD J. OF LEGAL STUD. 899, 899 (2021).

[106] *See generally* DANIEL KAHNEMAN, ET AL., NOISE (2021).

[107] Rebecca Heilweil, *Facebook is Taking a Hard Look at Racial Bias in its Algorithms*, VOX (22 July 2020), https://www.vox.com/recode/2020/7/22/21334051/facebook-news-feed-instagram-algorithm-racial-bias-civil-rights-audit.

[108] *See generally* Barrie Sander, *Democratic Disruption in the Age of Social Media: Between Marketized and Structural Conceptions of Human Rights Law*, 32 EUR. J. INT'L L. 159, 166-173 (2021).

censorship, users could still receive information based in fact that vast amounts of misinformation might otherwise drown out within their informational bubbles.[109]

Upon reflecting on the amount of misinformation that can appear on social media at certain times, a consideration that becomes clearer is *when* to use the alternative source digital nudge. Although this appears not to have been publicly discussed, one proposal worth setting out is that the time for a social media platform to implement alternative source digital nudging would be when there are spikes in misinformation on that platform. While this digital nudge could operate continually on social media platforms, its use could also be limited to periods of time where it would arguably become more necessary, due to the specifics of what is at stake during a particular influx of misinformation.[110] As such, social media platforms might consider utilizing the alternative source digital nudge in the lead up to elections; during armed conflicts, disasters, and pandemics; and after attacks that may be labeled as terrorism, where members of communities can suffer from falsehoods being ascribed to their group.[111]

The alternative source digital nudge has the potential to facilitate the spread of accurate information without social media users themselves having to make the effort of seeking out new sources of information. Such sources would instead be neatly presented to them, ready to consume. This is significant because many if not most people are busy (some too busy), encaged within the broken neoliberal hamster wheels that are the default operating systems of many societies at present. There can be condescension attached to judgments toward people that may be uninformed or misinformed on a subject due to them not seeking out new information, by people that have the luxury of time to inform—and misinform—themselves by seeking out new information. Time is limited. And many people do not have the privilege of being able to consume new information throughout their day. Instead, unrelenting systems of oppression keep them too busy to do so.

---

[109] *Id.* at 164.

[110] *See generally* Erik C. Nisbet, Chloe Mortenson & Qin Li, T*he Presumed Influence of Election Misinformation on Others Reduces Our Own Satisfaction with Democracy*, 1 HARV. KENNEDY SCH. MISINFO. REV., 1 (2021) (discussing the impact of political misinformation); Sarah Brown, *MIT Sloan Research about Social Media, Misinformation, and Elections*, MIT: IDEAS MADE TO MATTER (Oct. 5, 2020), https://mitsloan.mit.edu/ideas-made-to-matter/mit-sloan-research-about-social-media-misinformation-and-elections.

[111] *See generally* Caroline Mala Corbin, *Terrorists Are Always Muslim but Never White: At the Intersection of Critical Race Theory and Propaganda*, 86 FORDHAM L. REV. 455 (2017); Alexandre Bovet & Hermán A. Makse, *'Influence of Fake News in Twitter during the 2016 US Presidential Election* 10 NATURE COMMC'N, 1 (2019).

Scrolling through a social media newsfeed is close to effortless, taking less time than seeking out and examining alternative sources of information and weighing them against prior knowledge. Such reasoning has been summarized elsewhere:

> Assessing the accuracy of information can be a difficult task. In today's fast-paced society, mass communication and social media play key roles in the sharing and receiving of current events. In reality, people generally do not have time to investigate each claim they encounter in depth; therefore, providing quality information is essential.[112]

Much responsibility can be placed on individuals to inform themselves on issues that concern them and their respective communities. However, doing so requires the time to make this effort. And the systems that create the conditions for this lack of time mean those that benefit from these systems—who also construct and maintain them—should exercise the responsibility that they sometimes attempt to pass on to individuals by way of "responsibilization."[113] Responsibilization refers to the transfer of responsibility away from actors with more power to address a particular issue to actors with less power to do so. The companies that own social media platforms have more power than any individual to reduce the spread of misinformation. So do states. It is arguably past time that these actors exercised responsibility in proportionate measure to such power. Digital nudging is not a panacea for misinformation, but these behavioral interventions show potential in helping manage it. They may also be more desirable than alternatives, even in light of the risks and shortcomings associated with their use.

C. Alternatives to Digital Nudging

This section will explore litigation, user-reporting, algorithmic downgrading, and content removal and deplatforming as alternatives to managing misinformation on social media platforms, compared to the two digital nudges analyzed above. These alternatives to digital nudging require attention, not only for deciding what measures are preferable in addressing misinformation, but also because prior research alludes to each one apparently being helpful in this respect (which may not be the case).

---

[112] Swire & Ecker, *supra* note 104, at 206.

[113] *See* Jarko Pyysiäinen, Darren Halpin & Andrew Guilfoyle, *Neoliberal Governance and 'Responsibilization' of Agents: Reassessing the Mechanisms of Responsibility-Shift in Neoliberal Discursive Environments*, 18 J. SOC. THEORY 215 (2017); Nikolas Rose, *Governing "Advanced" Liberal Democracies*, *in* FOUCAULT AND POLITICAL REASON: LIBERALISM, NEO-LIBERALISM, AND RATIONALITIES OF GOVERNMENT 37 (Andrew Barry, Thomas Osborne & Nikolas Rose eds., 1996).

Further reflection reveals the specific ways each measure raises governance challenges, including concerns regarding freedom of expression, information, and thought. This discussion also helps grasp in what ways these measures differ, while highlighting some commonalities.

1. Litigation: A Possible Vehicle for Change that Lacks Real-Time Practicality

Perhaps the most significant feature of digital nudging on social media newsfeeds is that it forms part of upstream governance.[114] By adopting this measure, the potential exists to address misinformation in real-time. This is in stark contrast to measures based on establishing liability, which form part of downstream governance, whereby the effects of misinformation would be judged after the fact. It is debatable that laws aiming to regulate misinformation by looking to accessorial or intermediary liability to function as a deterrent will change the associated practices of social media platforms.[115] This is especially debatable because of the difficulties in determining such liability, particularly establishing the nexus between the content containing misinformation and the resulting harm, and the extensive resources that owning companies have, which allow them to draw out lawsuits. Years of litigation over misinformation liability is not capable of reducing the spread of misinformation that has already occurred, even though any related ruling might result in platforms changing aspects of their modes of operation.[116] But bringing about such change also rests to an extent on suits not being settled privately and making it to trial, even though some settlement agreements may stipulate required changes to operating practices.[117] There is also the issue of whether companies will ultimately comply with the decisions reached after appeals processes have been exhausted—much of what social media platforms do is unknown to

[114] For an understanding of "upstream" and "downstream" governance, *see* Richard Mackenzie-Gray Scott, *Rebalancing Upstream and Downstream Scrutiny of Government During National Emergencies*, U.K. CONST. L. ASS'N (Sept. 21, 2021), https://ukconstitutionallaw.org/2021/09/21/richard-mackenzie-gray-scott-rebalancing-upstream-and-downstream-scrutiny-of-government-during-national-emergencies/.

[115] *But 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, 327 (2021) ("The expansion of intermediary liability, on the other hand, is likely to generate incentives for online media service providers to censor a greater amount of content for efficient identification of fake news").

[116] For a developing case, *see* Ian Millhiser, *Two GOP Judges Just Stripped Social Media Companies of Basic First Amendment Rights*, VOX (May 12, 2022, 3:00 PM), https://www.vox.com/2022/5/12/23068017/supreme-court-first-amendment-twitter-facebook-youtube-instagram-netchoice-paxton-texas.

[117] With thanks to Taylor Desgrosseilliers for raising this point.

outsiders, making it difficult to assess whether changes have been made without the details being made public.

Creating new laws (or getting creative with existing ones) in order to lessen protections for social media platforms against liability for the content contained on them is difficult. As Dorit Rubinstein Reiss notes:

> For example, if Facebook put out antivaccine misinformation and someone got hurt, it could be liable. That is a little tricky in two ways. First, it can really limit what social media companies can do. Facebook has over a billion users. Regulating all of those users is probably not feasible in real-time. There are going to be limits to what Facebook can do. Also, civil liability may mean that Facebook would completely shut down scenarios of discussion. That is not necessarily a good thing. Second, we are penalizing Facebook for things where other people are more culpable. We are not cutting the promoters of misinformation out.[118]

The efficacy of liability-based deterrents (those based on negative incentives) at influencing the conduct of individuals may be a sound premise even if at times ineffective, but with respect to collectives it is questionable.[119]

To the extent that litigation can only redress the outcomes of behavior and not influence behavior itself, digital nudging is a more cost-effective measure to manage misinformation on social media than lawsuits. For those that consider expected liability a deterrent to social media platforms' conduct, it is worth reflecting on the extent of these platforms' existing unlawful and allegedly unlawful practices, which may continue without interruption, even in the face of litigation.[120] There are things money cannot buy,[121] but a team of lawyers and legal strategists is not one of them.[122]

---

[118] Dorit Rubenstein Reiss, *Anti-Vaccine Misinformation and the Law: Challenges and Pitfall*s, 18 IND. HEALTH L. REV. 85, 92 (2020).

[119] *See, e.g.*, Jonathan Klick & John MacDonald, *Deterrence and Liability for Intentional Torts*, 63 INT'L. REV. L. AND ECON. 1, 1-2 (2020); Paul H. Robinson & John M. Darley, *Does Criminal Law Deter? A Behavioural Science Investigation*, 24 OXF. J. LEG. STUD 173, 173 (2004).

[120] *See, e.g.*, FACEBOOK CLAIM, https://www.facebookclaim.co.uk/; *FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook*, FTC (Jul. 24, 2019), https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions; Diane Bartz, *Texas Sues Meta's Facebook over Facial-Recognition Practices*, REUTERS (Feb. 14, 2022), https://www.reuters.com/technology/texas-sues-meta-over-facebooks-facial-recognition-practices-report-2022-02-14/; Kari Paul, *'Live in the Future': Zuckerberg Unveils Company Overhaul amid Shift to Metaverse*, THE GUARDIAN (Feb. 16, 2022), https://www.theguardian.com/technology/2022/feb/15/meta-mark-zuckerberg-facebook-metaverse; *Lawsuits Involving Meta Platforms*, WIKIPEDIA, https://en.wikipedia.org/wiki/Lawsuits_involving_Meta_Platforms (last visited Sep. 8, 2022).

[121] *See generally* MICHAEL J. SANDEL, WHAT MONEY CAN'T BUY: THE MORAL LIMITS OF MARKETS, (2013).

When the benefits of compliance with the law are outweighed by the benefits of non-compliance, legal rules and their loopholes become part of the problem in enabling practices that are unlawful for those able to afford the related legal challenges. A further factor to consider is that, with states attempting to influence companies to implement self-regulatory efforts in a manner geared toward avoiding potential liability for content on their social media platforms, regression in human rights safeguarding could occur. There is a risk that companies introduce sweeping content moderation decisions in response to any such state measures, which may negatively impact freedom of expression and access to information,[123] both of which are important for protecting and promoting freedom of thought.

2. User-Reporting: Aggregation Has Its Limits

The next alternative is user-reporting. This consists of providing users the opportunity to send private notifications to social media platforms when they believe a particular source contains misinformation, or write public notes attached to posts that call into question their content. In 2021 Twitter launched Birdwatch, a pilot flagging program aimed at helping "address misleading information on Twitter," which consists of allowing users to "provide informative context" accompanying posts that may eventually become "visible directly on Tweets for the global Twitter audience, when there is consensus from a broad and diverse set of contributors."[124] The logic behind this initiative is that aggregating individuals' views can produce more accuracy than any one person (such as a fact-checker), meaning scaling up practices focused on the identification of misinformation is promising.[125] Although this approach to identifying misinformation has intuitive appeal, especially considering aggregating a diverse set of numerous opinions can counteract confirmation bias, echo

[122] *See generally* KATHARINA PISTOR, THE CODE OF CAPITAL: HOW THE LAW CREATES WEALTH AND INEQUALITY (2019).

[123] *See* U.N. Secretariat, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶¶ 12-21, U.N. Doc. A/HRC/38/35 (April 6, 2018); U.N. Secretary-General, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 33, U.N. Doc. A/72/350 (Aug. 18, 2017).

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

[125] *See* Jennifer Allen, et al., *Scaling Up Fact-Checking Using the Wisdom of Crowds*, 7 SCI. ADVANCES 1, 6 (2021).

chambers, groupthink, and motivated reasoning,[126] it has a possible hitch: belief management at scale.

If social media platforms receive mountains of misinformation reports from users, it is questionable whether such submissions will be actionable in a timely manner. And if reports are publicly available, then there is the problem of large numbers of people reaching consensus on content that is misinformation, as consensus can be reached on anything, even if it is nowhere close to being true.[127] There is also the risk that user-reporting generates mass reporting of specific sources of information, meaning this measure could be taken advantage of for the purpose—or at least with the effect—of undermining the credibility of accurate sources.[128] In order to avoid propagating misinformation when the convergence of multiple opinions occurs, it will ultimately fall on social media platforms to decide when a point of consensus has been reached and its proximity to truth, which would be based on a particular platform drawing aggregated conclusions from samples of individual opinions.

Does this mean the aggregation of ideas can provide the closest picture of the truth? Not always. The Zollman effect says that less communication occurring between sources of information will aid in at least one source of information gathering evidence that ultimately confirms the truth in question, with the admonition that in "tightly connected networks, misleading evidence is widely shared, and may cause the community to pre-emptively settle on a poor theory."[129]

This arbiter of truth problem is also shared with fact-check alerts, although their means of operation are different. Digital nudging in the form of fact-check notifications relies on company resources to check facts, whether by doing so itself or utilizing the services of a third party. User-reporting essentially outsources this review work to the masses, where thousands if not millions of people will be working (more) for no pay to serve the interests of corporations. User-reporting can thus operate a bit like the peer-review system in many academic journals, which can be valuable if conducted properly, but arguably takes advantage of the generous efforts of editors and reviewers. Yet the edge digital nudges may

---

[126] *See* CASS R. SUNSTEIN & REID HASTIE, WISER: GETTING BEYOND GROUPTHINK TO MAKE GROUPS SMARTER 109-144 (2014).

[127] *See* CAILIN O'CONNOR & JAMES OWEN WEATHERALL, *How False Beliefs Spread*, *in* THE MISINFORMATION AGE 46, 46-92 (2019).

[128] With thanks to Jordan Plummer for raising this concern.

[129] *See* Alvin Goldman & Cailin O'Connor, *Social Epistemology*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2021), https://plato.stanford.edu/entries/epistemology-social/.

have over user-reporting is their ability to apply in real-time, because aggregating hundreds, thousands, or millions of diverse opinions may not be capable of quickly countering rapidly spreading misinformation.

That said, social media platforms possess immense computing power, which means rapid aggregation may be possible — depending on multiple user reports being contributed in a short timeframe. Perhaps the most notable feature that user-reporting has over digital nudges is that it provides users with more agency in the process of managing misinformation. People are invited to become active participants in this effort, collaborating with companies, instead of being passive subjects of behavioral interventions imposed by companies.

Yet there is another factor associated with user-reporting that calls for attention: who collaborates with social media companies, and why? As part of the user-reporting process, a state may issue a request to remove content that it considers to be misinformation. However, the content may not actually be misinformation, but, for example, a critique of the state. The dangers of such circumstances are manifold. In order to maintain access to the market in a particular state, a social media platform might comply with a state request to withhold content on the grounds that it allegedly contains misinformation.[130] The compliance incentive here concerns the threat of adopting state policies that can potentially reduce or eliminate revenue streams associated with that market should the platform not comply with the request in question. Cases of this sort present a mutual benefit to the two actors—states can suppress dissident voices, and companies can maintain access to revenue. The more the mutual benefit intersection increases between a state and a social media platform, the greater the potential risk for certain sources to be censored when they disagree with a particular state position, including when accounting for the switching cost reasoning of users.[131] This potential problem is, at least logically, held in common with digital nudges, because both content that is labeled misinformation by fact-check alerts, and content presented as an alternative source, may raise objections from states seeking to suppress certain sources by accusing them of promulgating misinformation.

---

[130] *See, e.g.*, *About country withheld content*, TWITTER (last visited Fed. 2, 2023), https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country (outlining the Twitter Platform Use Guidelines: "if we receive a valid and properly scoped request from an authorized entity, it may be necessary to withhold access to certain content in a particular country from time to time. Such withholdings will be limited to the specific jurisdiction that has issued the valid legal demand or where the content has been found to violate local law(s)").

[131] Shin-Ru Cheng, *Market Power and Switching Costs: An Empirical Study of Online Networking Market*, 90 UNIV. CIN. L. REV. 122 (2021).

3. Algorithmic Downgrading: Lacking in Transparency and Depriving Agency

Algorithmic downgrading consists of an algorithm demoting content posted on social media so that it appears further down on users' newsfeeds and is used to control information that users receive. On platforms that use downgrading in an attempt to reduce the spread of misinformation, users are less likely to notice content containing misinformation compared to content that appears further up in their newsfeeds.[132] This practice is about as troublesome as it is opaque. It robs users of agency and relies on hidden mechanisms informed by unknown factors.

Social media platforms rank content before presenting it to users on their newsfeeds with the aim of ensuring that people keep scrolling, clicking, and typing for the longest possible periods of time, thereby divulging more data about themselves and facilitating further tailoring of their experience, leading to more time spent on the platform.[133] Like PageRank, the algorithm used for Google Search, algorithmic design focused on ranking content so that it is customized to maximize user engagement can "decrease the diversity of news sources that people see."[134] The continuation of this mode of operation has potentially contributed to a filter bubble[135] where many sources of information remain hidden from users' view because they do not make it through an algorithm's filter.[136]

In light of this situation, lawmakers are debating whether to alter how social media platforms curate what content users perceive on their newsfeeds. In the U.S., Senators Amy Klobuchar and Cynthia Lummis introduced the Social Media Nudge Act in February 2022.[137] This legislation aims to require "the Federal Trade Commission to identify

---

[132] Will Oremus, *Why Facebook Won't Let You Control Your Own News Feed*, *The Washington Post*, WASH. POST (Nov. 15, 2021), https://www.washingtonpost.com/technology/2021/11/13/facebook-news-feed-algorithm-how-to-turn-it-off/.

[133] *See* Terry Flew & Petros Iosifidis, *Populism, Globalisation and Social Media*, 82 INT'L COMMC'N GAZETTE 7, 19-20 (2020).

[134] Petros Iosifidis & Leighton Andrews, *Regulating the Internet Intermediaries in a Post-Truth World: Beyond Media Policy?*, 82 INT'L COMMC'N GAZETTE 211, 218 (2019).

[135] *See* Amy R. Arguedas, et al., Echo *Chambers, Filter Bubbles, and Polarisation: A Literature Review*, REUTERS INST. FOR STUD. JOURNALISM, 10-11 (2022), https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review; *see also Our funders*, REUTERS INST. FOR STUD. JOURNALISM (last visited Feb. 3, 2023), https://reutersinstitute.politics.ox.ac.uk/our-funders (noting that the Reuters Institute for the Study of Journalism at the University of Oxford is funded by the Google News Initiative and the Facebook Journalism Project, among other funders).

[136] Iosifidis & Andrews, *supra* note 135; *see also* ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK 16 (2011).

[137] Social Media NUDGE Act, S. 3608, 117th Cong. (as read twice and referred to the Committee on Commerce, Science, and Transportation, Feb. 9, 2022).

content-agnostic platform interventions to reduce the harm of algorithmic amplification and social media addiction on covered platforms, and for other purposes."[138]

Whether enacted or not, algorithmic downgrading (or indeed upgrading or promoting content) based on ranking sources of information against multiple metrics goes unseen. While the outcomes of implementing this measure may be considered desirable (for example, demoting bogus claims about personal health or encouraging vaccine uptake),[139] that its process remains hidden from users, as well as everyone other than employees of social media companies that are privy to the related details, is concerning. Such opacity "is disempowering for users and denies agency."[140]

Transparency is a component of effectively addressing misinformation because of its link to trust. It can be hard to trust the outcome of a process if there is no way of knowing what that process entails.[141] In comparison to algorithmic downgrading based on the ranking of information against unknown metrics, digital nudges are more transparent. Whereas users do not know whether algorithmic downgrading occurs on their newsfeeds, digital nudges are literally presented in front of users. The key is recognizing a digital nudge when it is presented. Acquiring the knowledge necessary for such recognition means educational efforts regarding nudging are linked to the extent to which digital nudges can be a transparent measure for addressing misinformation. Further education of this sort may also help alleviate the criticisms about nudges being manipulative interventions, because once a person understands what something is and recognizes its purpose, it can hold less influence over that person.[142] Digital

---

[138] *Id*.; See Shirin Ali, *Congress Might Try to Force Facebook to Change its Newsfeed Algorithm*, THE HILL (Feb. 11, 2022), https://thehill.com/changing-america/well-being/mental-health/593852-congress-might-force-facebook-to-change-its.

[139] Renee Garett & Sean D. Young, *Online Misinformation and Vaccine Hesitancy*, 11 *Translational Behavioral Medicine* 2194 (2021); Andrew Hutchinson, *Facebook Updates News Feed Algorithm to Demote Misleading Health Claims* SOCIAL MEDIA TODAY (Jul. 03, 2019), https://www.socialmediatoday.com/news/facebook-updates-news-feed-algorithm-to-demote-misleading-health-claims/558100/.

[140] Irene Khan (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Disinformation and Freedom of Opinion and Expression*, ¶ 80, U.N. Doc. A/HRC/47/25 (Apr. 13, 2021).

[141] Heungsik Park & John Blenkinsopp, *The Roles of Transparency and Trust in the Relationship between Corruption and Citizen Satisfaction*, 77 INT'L. REV. OF ADMIN. SCI. 254, 257-8 (2011).

[142] Thomas R. V. Nys & Bart Engelen, J*udging Nudging: Answering the Manipulation Objection*, 65 POL. STUDIES 199 (2017); Martin Wilkinson, *Nudges Manipulate, Except When They Don't*, LONDON SCH. OF ECON., BRITISH POL. AND POL'Y. (Aug. 29, 2013) https://blogs.lse.ac.uk/politicsandpolicy/nudges-manipulate-except-when-they-dont/.

nudges also leave room for users to choose what information they consume, as opposed to being kept away from accessing certain sources of information. The further a source of information is downgraded on a newsfeed, the harder it becomes for a user to access it.

If designed properly, the alternative source digital nudge may actually be capable of counteracting any filter bubble effect of ranking algorithms. In addition to information that is filtered out of users' newsfeeds, similar or even the same information could be reintroduced onto the platform, albeit not as original content, but instead accompanying such content as an alternative source. The upgrading and downgrading of sources would thus have an additional component: presenting new information that would not normally be made available to a user on a particular platform. Perhaps tweaking newsfeed algorithms in this way is a middle ground between the companies owning social media platforms and the states aiming to regulate their newsfeeds. Ranking algorithms could still be used similarly to how they are already, but in a way in which content that would ordinarily be completely filtered out of a newsfeed is instead introduced via an alternative source digital nudge.

Even if this idea is considered problematic or unworkable, the use of digital nudges still grants more agency to social media users in managing misinformation than algorithmic downgrading. In comparative terms, users would be brought further into the loop in deciding what information to consume, instead of being cut out of this decision by the preferences set by companies and their subsequent crafting of automated systems of information filtration.

4. Content Removal and Deplatforming: Mirages of Human Rights Marketing?

Questions of individual autonomy also underlie the contrast between digital nudging and two additional measures of combating misinformation on social media platforms: content removal and deplatforming. Resorting to these measures has attracted attention, both in the form of disapproval and praise.[143] They give the appearance of being an effective method to reduce the spread of misinformation. These measures remove sources of misinformation on a platform when a social media company decides that a particular threshold has been met. Skipping over the matter of the

[143] Shirin Ghaffary, *Does Banning Extremists Online Work? It Depends*, VOX (Feb. 03, 2022), https://www.vox.com/recode/22913046/deplatforming-extremists-ban-qanon-proud-boys-boogaloo-oathkeepers-three-percenters-trump; Clive Cookson, S*ocial media sites should not ban misleading content, UK scientists say*, FINANCIAL TIMES, Jan. 19, 2022, https://www.ft.com/content/9cf1ee59-985c-4a71-ac96-895cd6413703.

unknowns regarding what this threshold is in a particular context and how it is determined by a particular platform (another transparency issue), the Royal Society issued a report in 2022 recommending that "[g]overnments and social media platforms should not rely on content removal as a solution to online scientific misinformation."[144] In summary, it highlights that open discussion and information sharing is a good practice, and trying to prevent it contributes to distrust and can hinder progress in terms of truth-seeking, including because deciding what is misinformation and what is not is resource intensive.[145]

A footnote in this report also provides a reference to the Streisand effect.[146] This describes situations where attempts to suppress information instead cause the information to receive more attention than it might have otherwise attracted.[147] By removing content or deplatforming sources that share it, social media platforms could end up lending misinformation more credence than it might have received if it had been left to linger, especially considering the speed at which newsfeeds churn out fresh content. There is also the risk of content removal and deplatforming resulting in more people buying into conspiracy theories due to the fires of their curiosity being stoked by questions surrounding why a particular source of information was removed from a platform.

For example, Facebook deplatformed Donald Trump's account in 2021, looking "to experts to assess whether the risk to public safety has receded,"[148] and ultimately deciding to reinstate the account in February 2023.[149] Now that Trump has his account back on this and other platforms that also instituted a ban,[150] if he should run for the U.S. presidency again,[151]

---

[144] THE ROYAL SOCIETY, THE ONLINE INFORMATION ENVIRONMENT: UNDERSTANDING HOW THE INTERNET SHAPES PEOPLE'S ENGAGEMENT WITH SPECIFIC INFORMATION 10 (2022), https://royalsociety.org/-/media/policy/projects/online-information-environment/the-online-information-environment.pdf

[145] *Id*. at 10-11, 49-50, 62-69.

[146] See *Id.* at 11 n.34.

[147] *See* Kieran Andrews, *Scottish Election: Pro-Scottish Independence Blog from Westminster Trade Adviser Deleted,* THE TIMES (Apr. 3, 2021), https://www.thetimes.co.uk/article/scottish-election-pro-scottish-independence-blog-from-westminster-trade-adviser-deleted-3gc9gmncn; The National Newsdesk, *Westminster refuses to deny it pushed academics to delete blog on indy Scotland*, THE NATIONAL (Apr. 2, 2021) https://www.thenational.scot/news/19208021.westmi.

[148] Nick Clegg, *In Response to Oversight Board, Trump Suspended for Two Years; Will Only Be Reinstated if Conditions Permit*, FACEBOOK (Jun. 4, 2021), https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/.

[149] Phelan Chatterjee, *Donald Trump Back on Facebook and Instagram*, BBC (Feb. 9, 2023), https://www.bbc.com/news/technology-64585429.

[150] James Clayton, *A Year On, Has Trump Benefited from a Twitter Ban?*, BBC (Jan. 12, 2022), https://www.bbc.com/news/technology-59948946.

will the spread of misinformation through his followers' networks have been reduced, increased, or remained roughly the same?[152]

Although decisions to deplatform accounts or remove content may have a short-term impact on reducing the spread of misinformation, in the long-term this is unproven.[153] One factor that is worth accounting for is that people can migrate to other social media platforms.[154] The switching cost for users is reduced when sources are deplatformed and content is removed by one platform, because lack of access on that platform makes moving to another an easier decision. This emphasizes an incentive for social media companies to not resort to content removal and deplatforming as doing so means losing users.[155] Further, even short-term reductions in the spread of misinformation may be overstated or erroneous considering proxy accounts can still distribute the same messages as original sources, albeit with less of a following.[156]

Facebook's deplatforming of Donald Trump is also significant for another reason. While some believe its Oversight Board made the "right call" in upholding Facebook's decision,[157] including because the members of this body referred to human rights law in their reasoning,[158] it is unclear

---

[151] David Frum, *Revenge of the Donald*, THE ATLANTIC (Oct. 28, 2021), https://www.theatlantic.com/ideas/archive/2021/10/trump-running-president-2024-election/620502/.

[152] *See* Gabby Orr, Kristen Holmes & Veronica Stracqualursi, *Former President Donald Trump Announces a White House Bid for 2024*, CNN (Nov. 16, 2022), https://edition.cnn.com/2022/11/15/politics/trump-2024-presidential-bid/index.html.

[153] Brad Honigberg, *Why Deplatforming Just Isn't Enough*, CTR. FOR STRATEGIC & INT'L STUD. (Feb. 11, 2021), https://www.csis.org/blogs/technology-policy-blog/why-deplatforming-just-*isnt*-enough; Eileen Guo, *Deplatforming Trump will work, even if it won't solve everything*, MIT TECH. REV. (Jan. 8, 2021), https://www.technologyreview.com/2021/01/08/1015956/twitter-bans-trump-deplatforming/.

[154] Richard Rogers, *Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media*, 35 EUR. J. OF COMMC'N 213, 214-15 (2020).

[155] Cory Doctorow, *Facebook's Secret War on Switching Costs*, ELEC. FRONTIER FOUND. (Aug. 27, 2021), https://www.eff.org/deeplinks/2021/08/facebooks-secret-war-switching-costs.

[156] Being unable to verify who actually uses a social media account is also a problem – even blue ticks next to friendly-looking faces does not necessarily mean the owners of those faces operate the accounts showcasing them and their supposed viewpoints. *See* Reuters in São Paulo, *Brazil Military 'Posed as NGOs on Social Media' to Play Down Deforestation*, THE GUARDIAN (Apr. 7, 2022), https://www.theguardian.com/technology/2022/apr/07/brazil-military-posed-as-ngos-on-social-media-to-play-down-deforestation.

[157] Marko Milanovic, *The Facebook Oversight Board Made the Right Call on the Trump Suspension*, EJIL: TALK! (May 6, 2021), https://www.ejiltalk.org/the-facebook-oversight-board-made-the-right-call-on-the-trump-suspension/.

[158] For the decision, *see* Oversight Board, Case decision 2021-001-FB-FBR (Jan. 7, 2021); For personal insights into the deliberations, *see* Alan Rusbridger, *The Inside Story of How We Reached the Facebook-Trump Verdict*, THE GUARDIAN (May 5, 2021), https://www.theguardian.com/commentisfree/2021/may/05/trump-facebook-oversight-board-verdict-alan-rusbridger.

how content curation and moderation is influenced by the work of the Oversight Board generally. David Morar takes the view that the process and outcome in this case "seems to be the product of a corporate pseudo-judiciary trying desperately to also be a corporate pseudo-legislative."[159] The Oversight Board's court-like appearance requires some explanation. Those that liken this body to a judiciary can be forgiven, as perhaps a key motivation behind Facebook creating it was to give the appearance that the company was establishing an accountability mechanism that might look like a duck and quack like a duck, but is not a duck. Courts can hold entities to account for their conduct, including Facebook. The Oversight Board cannot do so in the same way. While the public relations language of "overturning" Facebook decisions gives the impression that the company is legally obligated to comply with Oversight Board rulings, it is unclear whether and to what extent Facebook is so required, regardless of what the terms used in these rulings might convey.[160]

The Oversight Board provides a description of the mandate it was given by Facebook: "The board's decisions to uphold or reverse Facebook's content decisions will be binding, meaning Facebook will have to implement them, unless doing so could violate the law."[161] The definition of "binding" is not clear here. Furthermore, for every instance where a decision might be compliant with "the law" (and what law?), there exists a possibility of that decision being a violation where the law is unclear. The likelihood of this occurring depends in part on the clarity provided in the law at issue and on those interpreting it. And if it is laws enshrining human rights at issue, ambiguity is widespread at the domestic, regional, and international levels, including, as is shown below, with respect to freedom of thought.

These features of this governance framework allow Facebook to continue operating with considerable discretion in its content moderation practices. If Facebook removes or restores any content or source based on decisions of the Oversight Board, it is doing so by choice.[162] The Oversight

---

[159] David Morar, *Trump Deplatforming Decision Highlights the Impotence of Facebook's Oversight Board*, BROOKINGS INST. (May 7, 2021), https://www.brookings.edu/blog/techtank/2021/05/07/trump-deplatforming-decision-highlights-the-impotence-of-facebooks-oversight-board.

[160] See *Case Decisions and Policy Advisory Opinions*, OVERSIGHT BD., https://www.oversightboard.com/decision/ (last updated Jan. 17, 2023), for the Oversight Board decisions that have been reached so far.

[161] OVERSIGHT BD., https://www.oversightboard.com/ (last visited Feb. 8, 2023).

[162] *See* Evelyn Douek, *The Facebook Oversight Board's First Decisions: Ambitious, and Perhaps Impractical*, LAWFARE (Jan. 28, 2021, 11:23 AM), https://www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical.

Board also risks being a scapegoat for Facebook, a means of passing the buck when the company gets things wrong. Attempting to avoid making decisions on content moderation is also a potential product of this relationship, with both bodies possibly ending up in a stasis of abdication. The preemptive deflection that can occur has already been displayed:

> Facebook asked its Oversight Board to review its decision to indefinitely ban Donald Trump, and guide it on whether it should allow the former president to post again. You could see it as the ultimate buck-passing. For three years, Facebook has been setting up an elaborate structure for a supposedly independent body to review its content decisions. And now that the 20-member board has just begun to hear cases, Facebook outsourced it with perhaps the company's most controversial decision ever . . . But the board did not play. While affirming that Facebook was correct to suspend the Trump account for its riot-coddling posts on January 6, today it called out the company for inventing a penalty that wasn't part of its policies—an 'indefinite' suspension. The board told Facebook to take six months and get its own rules straight, and then make the Trump restoration decision itself.[163]

The Oversight Board has limited power to hold Facebook to account for its content curation and moderation practices beyond making statements that may gain traction in the public arena, and its primary source of authority to undertake its work is granted and limited by the very company it is tasked with scrutinizing. Despite being funded by an independent trust, some also question its independence from Facebook. For example, Members of Parliament in the UK asked how much members of the Oversight Board are being paid.[164]

Conflating the decision-making processes of the two entities is hazardous; it endangers what appears to be the genuine engagement with human rights law by the Oversight Board. Time may tell if Facebook intended this body to belong in the toolbox of human rights marketing, deployed as a distraction, instead of in the toolbox of human rights law, which can be used to protect human beings. A distinction to bear in mind is that even though the Oversight Board's decisions are transparent in their reasoning and use the language of human rights law, this does not mean Facebook does the same when deciding what content to remove or restore.

---

[163] Steven Levy, *Oversight Board to Facebook: We're Not Going to Do Your Dirty Work*, WIRED (May 5, 2021), https://www.wired.com/story/oversight-board-to-facebook-not-going-to-do-your-dirty-work.

[164] James Cook, *MPs Urge Facebook to Reveal Pay of Alan Rusbridger and Other Members of its "Supreme Court,"* THE TELEGRAPH (May 7, 2020), https://www.telegraph.co.uk/technology/2020/05/07/mps-urge-facebook-reveal-pay-alan-rusbridger-members-supreme.

Indeed, the Oversight Board has demanded more transparency from Facebook.[165]

Bringing these insights back to the comparison with digital nudges, it is clear that content removal and deplatforming raise many thorny questions of how to reduce misinformation and its spread through social media. These primarily concern freedom of expression, access to information, and the related debate, discourse, and dissent; all of which are essential for the maintenance of a healthy democracy, in part because they help dilute concentrations of misinformation. Donato Vese has argued that from the perspective of democracy, digital nudges on social media are preferable to other types of state intervention based on censorship.[166] There is also no apparent tension between digital nudging and the right to freedom of expression insofar as manifestation is concerned.[167] Whereas, as shown by the work of David Kaye and others,[168] there is plenty of tension with respect to the measures of content removal and deplatforming.[169] So much so that a company that is not presently bound by human rights law became concerned enough to create an apparatus in order to bolster its appearance of taking its human rights responsibilities seriously.[170] It matters that, in principle, the two digital nudges examined here raise fewer issues regarding the right to freedom of expression.

A possible exception may lie in states' corresponding positive legal obligation as it relates to the "freedom to hold opinions and to receive and impart information and ideas without interference."[171] Digital nudging can change how users receive and impart information on social media newsfeeds. Fact-check alerts alter how users receive information and alternative sources impart information differently compared to if they were presented as original sources. Although "without interference" concerns the

---

[165] Parmy Olson, *Don't Dismiss Facebook's Oversight Board. It's Making Some Progress*, BLOOMBERG (Oct. 25, 2021), https://www.bloomberg.com/opinion/articles/2021-10-25/facebook-s-oversight-board-is-the-only-lever-to-reform-the-social-media-behemoth#xj4y7vzkg.

[166] Donato Vese, *Governing Fake News: The Regulation of Social Media and the Right to Freedom of Expression in the Era of Emergency*, 13 EUR. J. RISK REGUL. 477, 503 (2022).

[167] *See generally* Dominic McGoldrick, *The Limits of Freedom of Expression on Facebook and Social Networking Sites: A UK Perspective*, 13 HUM. RTS. L. REV. 125 (2013).

[168] *See generally* DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET (2019).

[169] Robert Spano, *Intermediary Liability for Online User Comments under the European Convention on Human Rights*, 17 HUM. RIGHTS L. Rev. 665 (2017); Lisl Brunner, *The Liability of an Online Intermediary for Third Party Content: The Watchdog Becomes the Monitor: Intermediary Liability after Delfi v Estonia*, 16 HUM. RTS L. Rev. 163 (2016).

[170] *See* Levy, *supra* note 164.

[171] Human Rights Act (1998), Sch. 1, P. I, Art. 10 (UK).

conduct of a "public authority" regarding the state's negative legal obligations,[172] which raises different questions should a state implement digital nudges itself,[173] a reasonable argument could be made that the state also has a positive legal obligation to protect people from non-state actors that may interfere with how information is received and imparted through digital nudging. But how far could any such argument hold before crumbling in relativism and irrationality? If the state has a positive legal obligation to protect against any use of digital nudges by social media companies on newsfeeds because they somehow "interfere" with how information is received and imparted, it would have to extend this protection wherever applicable.

This means the state would also arguably need to address non-state actors that interfere with informational exchange, whether, for example, by putting journals behind a paywall, restricting access to about 45% of court judgments,[174] or otherwise only permitting access to information for a fee. It is also worth noting that when assessing whether a state should act on a positive legal obligation to do something with the aim of protecting corresponding rights-holders,[175] consideration is often given to whether that state can bear the related burden "without abandoning other responsibilities that ought not to be abandoned."[176]

While digital nudging may yet be exposed as an enemy of free expression, there is little indication at present that this measure interferes with the exercise and enjoyment of this right. Even if this were the case, the exercise of this freedom "carries with it duties and responsibilities," which make it subject to "formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society."[177] Should states consider digital nudging on newsfeeds to be in the public interest to an extent where there is political appetite to introduce related laws and policies, managing misinformation and reducing its spread could fall within the ambit of pursuing a legitimate aim. Considering the range of situations

---

[172] *Id*.

[173] *See* Sofia Ranchordas, *Nudging citizens through technology in smart cities* 34 INT'L REV. OF L., COMPUT. & TECH. 254, 254 (2020).

[174] Daniel Hoadley, Joe Tomlinson, Editha Nemsic, & Cassandra Somers-Joce, *How Public is Public Law? Approximately 55%*, UK CONSTITUTIONAL L. ASS'N (Feb. 25, 2022).

[175] Wesley Hohfeld, *Some Fundamental Legal Conceptions as Applied in Judicial Reasoning,* 23 YALE L. J. 16, 28-59 (2013).

[176] James W. Nickel, *How Human Rights Generate Duties to Protect and Provide*, 15 HUM. RTS. Q. 77, 81 (1993); *see also* STEPHEN HOLMES & CASS R. SUNSTEIN, THE COST OF RIGHTS: WHY LIBERTY DEPENDS ON TAXES (2000).

[177] Human Rights Act (1998), Sch. 1, P. I, Art. 10 (UK).

to which misinformation can contribute at scale, such as public disorder, crime, and ill health, such regulation may also be necessary in particular contexts. It may also prove more proportionate than removing content from social media platforms or deplatforming individual accounts. Public policy considerations of this sort may be part of the future if developments in the U.S. regarding the Social Media Nudge Act are an appropriate indicator.[178]

Yet it is crucial that free expression is not hindered by any state. Death of this freedom spells the Orwellian for a society.[179] The "protection of individual autonomy requires us to protect the right of all to express themselves freely and their right to have access to the thoughts and ideas of others. Protecting free speech is therefore seen as enabling the self-fulfillment of individual members of a society."[180] Perhaps digital nudges on social media newsfeeds can go even further than merely not being an unlawful interference with this right. The alternative source digital nudge in particular not only arguably respects this right of free expression—at least insofar as its manifestation is concerned—but can also promote it by presenting information that may otherwise go unseen by users on a particular social media platform. This is far from an interference with pluralistic informational exchange. However, the opposite is true when reflecting on the measures of content removal and deplatforming. They exist to "police the streets" of online information, and apparently not always very well. Digital nudges do something else: they help accurate information compete in an online attention market.

D. Appreciating the Risks and Shortcomings of Digital Nudging

With the possible exception of user-reporting, digital nudging is the least troublesome measure of those examined above. This measure is also in stark contrast to others, in that it can address misinformation sharing in real-time. Other measures may also interfere with the right to freedom of expression and the related exchanges of information that are part of the very truth-seeking efforts capable of dispelling misinformation. Empirically, there remains more to learn about whether fact-check alerts and alternative sources truly work at preventing misinformation from being shared further on social media platforms, even if it turns out that they

[178] Social Media Nudge Act, S.3608, 117th Cong. (2022).

[179] Marko Milanovic, *The Legal Death of Free Speech in Russia*, BLOG OF THE EUR. J. OF INT'L L. (Mar. 8, 2022), https://www.ejiltalk.org/the-legal-death-of-free-speech-in-russia/; GEORGE ORWELL, NINETEEN EIGHTY-FOUR (Penguin, 2008). GEORGE ORWELL, NINETEEN EIGHTY-FOUR (2008).

[180] Catherine O'Regan, *Hate Speech Online: an (Intractable) Contemporary Challenge?*, 71(1) CURRENT LEGAL PROBLEMS 403, 409 (2018).

reduce click-through rates on content containing misinformation. Applying both digital nudges may also prove more effective in this respect than relying on one alone.

Although the analysis and discussion thus far might convey an impression that digital nudges should be met with approval, there are a number of reasons to refrain from rolling out the red carpet. While these reasons are not exhaustive, they are provided for the purpose of striking a cautionary chord with proponents of digital nudging. Even if the practice of digital nudging on social media newsfeeds turns out to be compatible with the law on the human right to freedom of thought, these risks and shortcomings will remain.

The first risk is baking bias into any algorithms that are designed to introduce digital nudges onto newsfeeds. Diversity in the teams designing these behavioral interventions is therefore vital. However, the technology industry is renowned for its lack of diversity.[181] The culture of social media companies in particular also seems to involve a lot of guy-fiving for maintaining the industry-wide touchstone of "move fast and break things."[182] One of the beauties of diverse groups of people working together on a project is one catching what another misses. No one person knows everything, but many people can know and believe the same things. Including people that know different things, hold different beliefs, and think in different ways can counteract the imperfect knowledge and biases applicable to all human beings.

The people that are subjected to digital nudging also require a voice through representation and input during decision-making processes. The public's preferences matter, as do their views on measures that are designed to affect them. Nourishing deliberative democracy means people should have a say in whether or not digital nudges should be used on social media newsfeeds to manage misinformation. Discarding this notion accepts the continuance of companies doing what they please in their pursuit of more profit and power. There are mixed views on nudging.[183] Before digital

---

[181] Ian Bogost, *The Problem with Diversity in Computing*, THE ATLANTIC (Jun. 25, 2019), https://www.theatlantic.com/technology/archive/2019/06/tech-computers-are-bigger-problem-diversity/592456/.

[182] Greg Williams, *Silicon Valley's Culture of Breaking Things is Totally Broken*, WIRED (Jul. 6, 2018), https://www.wired.co.uk/article/move-fast-and-break-things-or-dont.

[183] *See, e.g.*, Jet G. Sanders, et al., *Lessons From the UK's Lockdown: Discourse on Behavioural Science in Times of COVID-19*, 12 FRONTIERS IN PSYCHOL. 1 (2021); Catharine Evers, David R. Marchiori, Astrid F. Junghans, Jolien Cremers, and Denise de Ridder, *Citizen Approval of Nudging Interventions Promoting Healthy Eating: the Role of Intrusiveness and Trustworthiness*, 18 BMC PUB.

nudges are rolled out further on social media platforms, the public should have opportunities to share their thoughts and have them acted upon by representatives that are accountable for their decisions. This might help bridge the gap between policy choices based on the advice of technocrats and the ideas that arise from members of the public pursuing the common good. Encouraging such civic virtue means providing platforms for people to be heard.

This is noteworthy for another reason. While their motives may be well-intentioned, those pushing the frontiers of technological innovation can fail to appreciate the risks of their creations until after the genie is out of the bottle. Some can also hold the troubling outlook that societal issues can all be resolved with tech-based approaches. Consider the technology (such as armed drones, surveillance systems, and vaxxports) that has been created and normalized in the name of addressing terrorism post-9/11 or the COVID-19 pandemic. The effectiveness of these technologies in achieving their supposed aims is suspect, independent of the multiple impacts they have on human beings.[184] The whiffs of consequentialism and utilitarianism that underlie the reasoning of many in favor of nudging can be caught by a similar trap, that being some rationalization justifying the use of a particular technological tool for the "greater good."[185] Such thinking often exerts at least a smidgen of the belief that those formally educated and credentialed know what is best for everyone. Somewhat unsurprisingly, this sometimes patronizing attitude of framing related policies as a means of trying to "help" "other" people make "smart" decisions that are "better" for them, rubs people the wrong way and inhibits trust.[186] The libertarian paternalism that the logic of digital nudges fits within is not immune from such criticism.[187] Those constructing the choice architecture for digital

HEALTH 1182 (2018); William Hagman, David Andersson, Daniel Västfjäll and Gustav Tinghög, *Public Views on Policies Involving Nudges*, 6 REV. OF PHIL. AND PSYCHOL. 439 (2015).

[184] Stefania Milan, et al., *Promises Made to Be Broken: Performance and Performativity in Digital Vaccine and Immunity Certification*, 12 EUR. J. OF RISK REGUL. 382 (2021); Alexandre de Figueiredo, Heidi J. Larson & Stephen D. Reicher, *The Potential Impact of Vaccine Passports on Inclination to Accept COVID-19 Vaccinations in the United Kingdom: Evidence from a Large Cross-sectional Survey and Modeling Study*, 40 ECLINICALMEDICINE 1 (2021); Marko Milanovic, *Human Rights Treaties and Foreign Surveillance: Privacy in the Digital Age*, 56 HARV. INT'L. L. J. 81 (2015).

[185] For a different account of utilitarian thought that includes human rights, *see, e.g.*, JOHN STUART MILL, ON LIBERTY (1865); *See also* JOHN STUART MILL, UTILITARIANISM (1879).

[186] *See, e.g.*, MICHAEL J. SANDEL, THE TYRANNY OF MERIT: WHAT'S BECOME OF THE COMMON GOOD? 81-112 (2020); Göran Adamson, *Why Do Right-Wing Populist Parties Prosper? Twenty-One Suggestions to the Anti-Racist*, 56 SOCIETY 47 (2019); Thomas Hanitzsch, Arjen Van Dalen & Nina Steindl, *Caught in the Nexus: A Comparative and Longitudinal Analysis of Public Trust in the Press*, 23 INT'L. J. OF PRESS/POL. 3 (2018).

[187] CASS R. SUNSTEIN, WHY NUDGE? THE POLITICS OF LIBERTARIAN PATERNALISM (2014).

nudges on social media newsfeeds are not empowering users to overcome their cognitive quirks that make them susceptible to giving misinformation credence and sharing sources of it. They are aiming to influence decision-making toward outcomes that the digital nudge designers consider best.

Although they assist in the good governance of a society, the principles of economic efficiency and optimization to which social media companies cling have attained close to untouchable levels of deference, including by judiciaries.[188] This is worrying because in creating digital tools that adhere to these principles, other principles can be disregarded or considered only as an afterthought. As more aspects of life become dehumanized through automation, the humanness of humans must not be lost when considering and enacting responses to societal challenges,[189] including digital nudging as a means of reducing misinformation on social media.[190] While human judgment is flawed (influenced by factors such as getting tired, overconfidence when in a position of power, or being intimidated by social status)[191] and inferior to that which is computational to the extent that a judgment can be numericized,[192] limiting judgment to these computerized confines is questionable.

Perhaps the most extreme (if unlikely) risk posed by digital nudges is something more sinister. If digital nudges are successful in changing behavior toward lending more credence to information that is closer to the objective truth, then what happens if, following this line of reasoning to its logical end, convergence occurs? Should such an eventuality transpire, multiple and previously divergent viewpoints may cease to exist. Instead, many people would hold the same or similar viewpoints. This progressive deradicalization problem is that which changes diverse opinions over time toward a centralized understanding, where social media users could be nudged toward this position through continuous, repeated, and steady exposure to information considered to be true by some, with the risk that it may not be true. Even if the sources presented as alternatives are different

---

[188] *See, e.g.*, The Motherhood Plan v. HM Treasury [2021] EWHC 309 (Admin), 67-85; For further commentary, *see* Richard Mackenzie-Gray Scott, *Judicial Scrutiny of COVID-19 Regulations in the UK: Addressing Deference to Data-Driven Decision-Making in Human Rights Cases* (Bingham Centre for the Rule of Law, Working Paper, 2021).

[189] Anuj Puri, *Moral Imitation: Can an Algorithm Really be Ethical?*, 48 RUTGERS L. REV. 1 (forthcoming 2020).

[190] For a related point *see* John Tasioulas, *The Role of the Arts and Humanities in Thinking about Artificial Intelligence (AI)*, ADA LOVELACE INST. (Jun. 14, 2021), https://www.adalovelaceinstitute.org/blog/role-arts-humanities-thinking-artificial-intelligence-ai/.

[191] Ulrik Franke, *First- and Second-Level Bias in Automated Decision-making*, 35 PHIL. & TECH. 1, 6 (2022).

[192] Kahneman, et al., *supra* note 77.

across users, if they are all aligned with the same understanding of what is true, then the digital nudges containing them would be presenting essentially the same stories, even if they were portrayed differently.

One problem with centralized understandings is that a popular perspective may not be true, or at least may be misleading. Much of this apprehension depends on the selection of sources: who is selecting them, how they are selected, and who decides who will make these decisions (and in what way).[193] Social media users that positively respond to alternative sources could ultimately close the feedback loops that inform the newsfeed algorithms that curate the content to be consumed, thereby being pushed closer toward a previously agreed-upon position on a particular subject.

Is this desirable? Although the problem of progressive deradicalization is a far-fetched risk of using digital nudging to manage misinformation on social media, it knocks loose an uncomfortable query about why and to what extent misinformation should be addressed, if doing so changes opinions across populations to the extent that they effectively become indistinguishable. If not designed with care, the impacts of fact-check alerts and alternative source digital nudging can dissuade people from sharing certain information and ideas. These measures can also potentially narrow the sources of information upon which users rely, thereby exhibiting traits toward regression in terms of access to information and, relatedly, freedom of thought.

Another broader (and less fanciful) consideration that further research on choice and freethinking has the potential to clarify is whether nudging can actually make good on a claim that is part of its construction: namely, that nudges maintain people's freedom to choose between options. Cass Sunstein asserts, "[t]o count as such, a nudge must fully preserve freedom of choice."[194] Yet what are considered to be nudges may not be nudges under this conceptualization, thus presenting what can be termed the nudge paradox: nudges promise to preserve the freedom to choose between options but may not be capable of delivering on this promise. If the way in which humans think is based on the separate cognitive mechanisms depicted as System 1 and System 2, the former being automatic and unconscious, and if nudges are intended to be used in a way that interfaces with this lackadaisical System 1, then how does this procedure preserve freedom of choice? Can choices even be made unconsciously, and, if so, are the resulting decisions really free?

---

[193] *See generally* Zuboff, *supra* note 34.

[194] Cass Sunstein, *The Ethics of Nudging*, 32 YALE J. REGUL. 413, 417 (2015).

Choice to an extent implies that decisions are freely made.[195] Without diving into the murky waters of the ongoing debate about whether humans have free will,[196] conscious choices are arguably a manifestation of the freedom to choose between options. This means people that switch to System 2 thinking when presented with a nudge would arguably be freely choosing whatever option they ultimately settled upon. But it is not evident that a behavioral response based on System 1 thinking would be a choice—at least not one made with any intelligibility.[197] This perspective also calls into question the so-called "as judged by themselves" standard,[198] because if judgments are being constructed by nudges without human awareness, "then *choice architects might be engineering the very judgment from which they are claiming authority.*"[199]

This hypothesis appears to be another reason why the whole premise behind deliberately designing choice architecture can be criticized for not leaving such things to chance.[200] While the two-system theory of thinking may yet be disproven, so long as it holds, interactions dependent on System 1 thinking in order to bring about changes in behavior appear to be less about choices and perhaps instead be something more akin to reaction manipulation. Nudging governance on this view would therefore be something in the ballpark of attempted thought recalibration that *relies on* mechanized human behavior, not that which promotes the ability of human beings to make informed, conscious decisions that are best for them. And what is best for one person is not necessarily best for another—a further faulty assumption of those seemingly eager to help others help themselves.[201]

---

[195] Kerri Smith, *Brain Makes Decisions Before You Even Know It*, NATURE (Apr. 11, 2008), https://www.nature.com/articles/news.2008.751.

[196] Bahar Gholipour, *Philosophers and Neuroscientists Join Forces to See Whether Science Can Solve the Mystery of Free Will*, SCIENCE (Mar 21, 2019), https://www.science.org/content/article/philosophers-and-neuroscientists-join-forces-see-whether-science-can-solve-mystery-free.

[197] Ben R. Newell & David R. Shanks, *Unconscious Influences on Decision Making: A Critical Review*, 37 BEHAV. AND BRAIN SCI. 1 (2014); Katie E. Garrison & Ian M. Handley, *Not Merely Experiential: Unconscious Thought Can Be Rational*, 8 FRONTIERS IN PSYCHOL. 1 (2017).

[198] Sunstein, *supra* note 195, at 429-433.

[199] *Id*. at 430-431 (emphasis original).

[200] *See generally* CHOICE ARCHITECTURE IN DEMOCRACIES: EXPLORING THE LEGITIMACY OF NUDGING (Alexandra Kemmerer et al. eds., 2017); *See also* T. Martin Wilkinson, *Nudging and Manipulation*, 61 POL. STUD. 341 (2013); Martin Lodge & Kai Wegrich, *The Rationality Paradox of Nudge: Rational Tools of Government in a World of Bounded Rationality*, 38 L. & POL'Y. 250 (2016).

[201] *See* Kirsty Mackay, *The Glasgow Effect: Examining the City's Life Expectancy Gap – a Photo Essay*, THE GUARDIAN (Feb. 26, 2021, 2:00 PM), https://www.theguardian.com/artanddesign/2021/feb/26/the-glasgow-effect-examining-the-citys-life-expectancy-gap-a-photo-essay (For example, the default

Any nudge is little more than an intervention that, although exhibiting the potential to alter behavior depending on the domain, can only address the symptoms of some societal issues. Digital nudges may be received as a reactive content moderation effort, which does not address the business models that underpin "the drivers of disinformation and misinformation."[202] Digital nudges on newsfeeds cannot treat the causes of misinformation and its spread. Initiatives aimed at treating these causes will require comparatively more concerted efforts and reforms, such as furthering education in digital literacy and making the subject of misinformation part of curricula in schools.[203] As Jeremy Waldron once highlighted, "I wish, though, that I could be made a better chooser rather than having someone on high take advantage (even for my own benefit) of my current thoughtlessness and my shabby intuitions."[204] This point is particularly noteworthy considering that, over time, overreliance on digital nudging to manage misinformation on social media newsfeeds could contribute to users subjected to this measure becoming "more dependent on decisional support" from the respective implementing platform(s).[205]

Improving the navigability of information circulating on social media is likely more appropriately and effectively achieved by enhancing people's ability to independently navigate online information environments, when compared to implementing measures that attempt to navigate people toward consuming pre-selected sources of information, or away from sources that a select group of people have deemed to be wrong in some way. It is important not to get lost in the allure of technological measures promising quick fixes to societal issues.[206] Misinformation on social media is not only a technical problem. It is a difficulty that is sociopsychological, requiring multipronged approaches that "target both the supply (for example, more efficient fact-checking and changes to platform algorithms and policies)

pension plan that counteracts present bias is not going to help the skint employee that kicks the bucket before being able to access it).

[202] Irene Khan (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), ¶ 65, U.N. DOC. A/HRC/47/25 (Apr. 13, 2021).

[203] Frances Yeoman & Kate Morris, *Why Media Education in Schools Needs to be About Much More than "Fake News"*, THE CONVERSATION (Jan. 7, 2020, 8:58 AM), https://theconversation.com/why-media-education-in-schools-needs-to-be-about-much-more-than-fake-news-129156.

[204] Jeremy Waldron, *It's All for Your Own Good*, THE N. Y. REV., (Oct. 9, 2014), https://www.nybooks.com/articles/2014/10/09/cass-sunstein-its-all-your-own-good/.

[205] *See* Christian Schubert, *Exploring the (Behavioural) Political Economy of Nudging*, 13 J. OF INSTITUTIONAL ECON. 499, 512-513 (2017).

[206] With many thanks to Halefom Abraha for helping me realize this point through an enjoyable discussion.

and the consumption (for example, accuracy nudges and enhanced media literacy) of misinformation."[207]

While social media platforms amplify misinformation's reach and impact, and digital nudges can be part of reducing this amplification, digital nudges are, at best, a complement to educational efforts that promote autonomy through learning. Funding of, and access to, such education is thus a related issue, meaning the systems of governance that prevent educational participation will also require treatment. The prominence of ideals that create the bureaucracy, price of existing, and lack of time for the majority of people to partake in their education is a considerable barrier to getting a handle on misinformation. Investment and its proper shepherding in education and related support structures is key. The nurse getting off a fourteen-hour shift and rushing to the nearest food bank for their family's first meal of the day is not going to make the new course on misinformation at the local college. Neither is the cabbie that works sixty hours a week on minimum wage in addition to undertaking unpaid care work. Economic models, social practices, and cultural norms that have become embedded with ideals that have led to inequality running rampant within and between states could well be a leading obstacle to addressing misinformation.[208]

Further debate is needed to reach a decision about whether digital nudges should be used on social media newsfeeds in order to manage the spread of misinformation. Whatever way the scales ultimately tip, the related contestations would do well to occur in public forums so as to provide for adequate scrutiny. Effective misinformation mitigation strategies also require an adaptation in outlook from companies that own social media platforms—one that balances their faith in utility maximization against human rights and the common good. While it is currently unclear if platforms' practices will realign in such a way, that they are making efforts to address misinformation is a step in a promising direction. Through empirical consensus on their efficacy or lack thereof, the use of digital nudges will become more or less justifiable on a number of grounds.

---

[207] Ullrich K. H. Ecker, et al., *The Psychological Drivers of Misinformation Belief and its Resistance to Correction*,1 NATURE REV. PSYCHOL. 1, 24 (2022).

[208] BRANKO MILANOVIC, THE HAVES AND THE HAVE-NOTS: A BRIEF AND IDIOSYNCRATIC HISTORY OF GLOBAL INEQUALITY (2011); Ravi Kanbur, *An Age of Rising Inequality? No, but Yes*, VOX EU CEPR (Sept. 21, 2020), https://voxeu.org/article/age-rising-inequality-no-yes; Thomas Goda & Alejandro Torres Garcia, *The Rising Tide of Absolute Global Income Inequality During 1850-2010: Is it Driven by Inequality Within or Between Countries?*, 130 SOC. INDICATORS RSCH. 1015 (2016).

But would such grounds include respecting the right to freedom of thought? Answering this question is undertaken with the aim of guiding debate and decision-making in government and industry on whether digital nudging on social media newsfeeds to manage misinformation should be further implemented, limited, or prohibited. Even if digital nudges become an unconvincing approach with respect to reducing misinformation in empirical terms, this does not necessarily mean social media platforms will put an end to their development and use, including potentially for other purposes. At present, fact-check alerts and alternative source presentation are unproven in behavioral terms and unsettling in moral ones, even if they are preferable to alternatives also aimed at addressing misinformation on social media. The next section speaks more to their lawfulness.

III. THE HUMAN RIGHT TO FREEDOM OF THOUGHT WHILE ENSNARED BY SOCIAL MEDIA NEWSFEEDS

There is no respite from the thundering river of content on social media newsfeeds. Once hooked up, users can struggle to pull themselves away from attention-seeking algorithms, bots, and humans. As there are many people used by these platforms, the implementation of digital nudges on them may prove to be effective at reducing the spread of misinformation at scale. Yet even if this is the case, whether this measure should be continued, expanded, limited, or discontinued can be informed by assessing its lawfulness, specifically regarding the right to freedom of thought. This is not to imply that other questions relating to the compatibility of digital nudges with other human rights laws are irrelevant or unimportant, quite the contrary, as the above discussion on freedom of expression revealed. The choice of focus here is because the right to freedom of thought has barely been analyzed in the context of using digital nudging on social media newsfeeds to manage misinformation,[209] even though the matter of nudging generally has been briefly noted in research that shares concerns about the extent to which accessing the digital realm is impacting the exercise and enjoyment of this right.[210]

---

[209] Richard Mackenzie-Gray Scott, *A Short-Term Option for Addressing Misinformation during Public Health Emergencies: Online Nudging and the Human Right to Freedom of Thought,* OPINIOJURIS (March 8, 2021), http://opiniojuris.org/2021/03/08/a-short-term-option-for-addressing-misinformation-during-public-health-emergencies-online-nudging-and-the-human-right-to-freedom-of-thought/, (last visited Feb. 2, 2023).

[210] Susie Alegre, *Regulating Around Freedom in the "Forum Internum,"* 21 ERA FORUM 591, 591 (2021); Simon McCarthy-Jones, *The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century*, 2 FRONTIERS ARTIFICIAL INTEL. 1, 1 (2019).

A few clarifications are necessary before delving into the details on whether digital nudging on newsfeeds is compatible with the right to freedom of thought as a matter of law. The purpose here is not to set out an argument that attempts to transpose the legal obligation held by states regarding the right to freedom of thought to social media companies. While the "state" is an abstract intersubjective construct of human creation, this legal fiction helps make sense of what laws apply to which actors and why. Nevertheless, it is important to acknowledge that states and non-state actors need not be different—they are just classified differently by those that wield the power (but not necessarily the authority) to do so.[211]

The seeming dichotomy between states and non-state actors is also false.[212] This includes with respect to the bearing of obligations under human rights law.[213] For example, there is evidence that non-state armed groups are bound by such legal obligations.[214] If true for armed groups, why not for other non-state actors, especially those that have an equal or even greater potential than any state to interfere with the exercise of human rights? The gradated approach of applying human rights law obligations to non-state actors conceptualized by Daragh Murray does so progressively,[215] meaning the further a non-state actor "displaces the power and authority of the state whose territory it is operating within,"[216] the stronger the case can

---

[211] *See* Janet McLean, *Problems of Translation: The State in Domestic and International Public Law and Beyond*, *in* THE FLUID STATE: INT'L L. & NAT'L LEGAL SYS. 210, 226 (Hilary Charlesworth et al. eds. 2005).

[212] *See* Sandesh Sivakumaran, *Beyond States and Non-State Actors: The Role of State-Empowered Entities in the Making and Shaping of International Law*, 55 COLUM. J. TRANSNAT'L L. 343, 343 (2017).

[213] Anthony Cullen & Steven Wheatley, *The Human Rights of Individuals in De Facto Regimes Under the European Convention on Human Rights*, 13 HUM. RTS. L. REV. 691, 691 (2013).

[214] S*ee, e.g.*, KATHARINE FORTIN, THE ACCOUNTABILITY OF ARMED GROUPS UNDER HUMAN RIGHTS LAW (2017); DARAGH MURRAY, HUMAN RIGHTS OBLIGATIONS OF NON-STATE ARMED GROUPS 2016); KONSTANTINOS MASTORODIMOS, ARMED NON-STATE ACTORS IN INTERNATIONAL HUMANITARIAN AND HUMAN RIGHTS LAW: FOUNDATION AND FRAMEWORK OF OBLIGATIONS, AND RULES ON ACCOUNTABILITY (2016); ANDREW CLAPHAM, HUMAN RIGHTS OBLIGATIONS OF NON-STATE ACTORS 271-317 (2006); N. S. Rodley, *Can Armed Opposition Groups Violate Human Rights?*, in HUMAN RIGHTS IN THE TWENTY-FIRST CENTURY: A GLOBAL CHALLENGE 297 (1993); Optional Protocol to the Convention on the Rights of the Child on the Involvement of Children in Armed Conflict (adopted 25 May 2000, entered into force 12 February 2002) A/RES/54/263, 2173 UNTS 222, Art. 4(1); African Union Convention for the Protection and Assistance of Internally Displaced Persons in Africa (Kampala Convention), Adopted by the Special Summit of the Union Held in Kampala, Uganda, 23 October 2009, Art. 2(e).

[215] *See* Daragh Murray, *Human Rights Obligations of Non-State Armed Groups*, EJIL: TALK! (Nov. 2, 2016), https://www.ejiltalk.org/book-discussion-introducing-daragh-murrays-human-rights-obligations-of-non-state-armed-groups-2/.

[216] Richard Mackenzie-Gray Scott, *State Responsibility for Complicity in the Internationally Wrongful Acts of Non-State Armed Groups*, 24 J. CONFLICT & SEC. L. 373, 404 (2019).

become for it bearing legal obligations. Yet, understandings of "territory" need not be limited to geographic locations that are offline.

In light of companies wielding significant power to negatively impact human beings, it is understandable that some advocate for transposing obligations under human rights law to these actors. Social media companies specifically also exercise a regulatory role over their online realms, more so than any state does at present.[217] In other words, social media companies function a bit like *de facto* regimes in the digital realm. All this being said, social media companies are not currently bound by human rights law,[218] despite the headway that the business and human rights movement has made in showing that companies have human rights responsibilities.[219] Therefore, when referring to the human right to freedom of thought, it is states that are the holders of the corresponding legal obligation. This means that if the use of digital nudges is found to be incompatible with this right, the positive legal obligation will fall on states to protect human beings from being subjected to their use by social media companies—thereby requiring the enactment of legislation limiting or prohibiting nudging on newsfeeds. Alternatively, other forms of state action would be needed to adequately provide such protection, such as public oversight committees requiring the submission of impact assessments.

In addition, an argument could be made that by failing to regulate the limitation or prohibition of any unlawful digital nudging on newsfeeds, social media companies' conduct in using these measures would be attributable to the states failing to regulate them appropriately. This would mean such states could be directly responsible for the use of this measure if it were contrary to the right of freedom of thought. Such an argument would be grounded in Article 9 of the International Law Commission Articles on State Responsibility, which reads:

> The conduct of a person or group of persons shall be considered an act of a State under international law if the person or group of persons is in fact exercising elements of the governmental authority in the absence or default of

---

[217] *See* I. Bremmer, *Big Tech Can See a Future Where the Nation State is No Longer the Master*, THE TIMES (Nov. 19, 2021), https://www.thetimes.co.uk/article/big-tech-can-see-a-future-where-the-nation-state-is-no-longer-the-master-ndt7cqzxf.

[218] *See, e.g.*, Molly K. Land, *Toward an International Law of the Internet*, 54 HARV. INT'L L. J. 393, 444–449 (2013); *See also* Richard A. Wilson & Molly K. Land, *Hate Speech on Social Media: Content Moderation in Context*, 52 CONN. L. REV. 1029, 1033 (2021).

[219] Office of the High Commissioner for Human Rights, Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework (2011), https://digitallibrary.un.org/record/720245?ln=en#record-files-collapse-header.

> the official authorities and in circumstances such as to call for the exercise of those elements of authority.[220]

The applicability of this provision rests on a situation of necessity existing. What this means regarding assessing attributability for the purposes of determining state responsibility in this context is that should a social media company attempt to regulate its newsfeed by implementing a measure (in this case digital nudging), which in the context of public administration is considered necessary (in this case because of an influx of misinformation), but is simultaneously contrary to a legal obligation held by a state (in this case regarding the right to freedom of thought), this conduct may be attributable to the state that failed to exercise governmental authority in this situation. The inverse would at least be an attempt by the state itself to address the misinformation. Instead, social media companies have had to step into this regulatory role and exercise such authority for the sake of the state and its people.

While it is more likely that states' positive legal obligation will be breached should the use of digital nudging be contrary to the right to freedom of thought, and accordingly states do nothing in response, it may be that the negative legal obligation can also be breached for such lack of action if attribution occurs. A key difference between these two approaches is the resulting responsibility (either direct or indirect), and the various consequences of the finding ultimately settled upon.[221] Yet the question this all hinges on is whether the implementation of digital nudging on newsfeeds to manage misinformation is compatible with the right to freedom of thought.

A. Freedom of Thought According to the Law and Its Interpreters

The significance of human rights law with respect to behavioral interventions on social media newsfeeds to manage misinformation is that its corpus of rules helps protect against the dangers posed by those in power that enact such measures. This includes decisions that can affect an entire population within a state and beyond when pursuing a collective

---

[220] Int'l L. Comm'n, Rep. on the Work of its Fifty-Third Session, U.N. DOC. A/56/10, at 49 (2001); G.A. RES. 56/83, ¶ 3 (Jan. 28, 2002); Rep. of the G.A., at 3, U.N DOC. A/CN.4/L.602/Rev.1 (2001); *Report of the International Law Commission to the General Assembly*, 56 U.N. GAOR Supp. No. 10, at 109, U.N. DOC. A/56/10.

[221] It should be noted that this dynamic is not necessarily dualistic. For further details *see* RICHARD MACKENZIE-GRAY SCOTT, STATE RESPONSIBILITY FOR NON-STATE ACTORS: PAST, PRESENT, AND PROSPECTS FOR THE FUTURE 156-164, 195-205, 246-247 (2022).

interest. Whether it is protecting public health or maintaining peace and security, powerholders can portray their policy choices as legitimate aims, and it is the role of human rights law to ensure that the conduct associated with such goals does not negatively impact the exercise and enjoyment of applicable human rights. As John Tasioulas puts it:

> Human rights law should secure human rights even if, as is often the case, doing so fails to maximise aggregate utility (e.g., because it rules out forms of surveillance or interrogation that would maximise general welfare but violate the right to privacy or the right not to be tortured).[222]

Even if digital nudging on newsfeeds becomes a clearly effective measure for managing misinformation on social media and reducing its spread, it should be abandoned if it is incompatible with the human right to freedom of thought.

Article 18 of the International Covenant on Civil and Political Rights (ICCPR) corresponds to the related provision under the Universal Declaration of Human Rights.[223] While requiring further incorporation at various national levels,[224] Article 18 sets out that "[e]veryone shall have the right to freedom of thought,"[225] albeit without further elaboration. The European Convention on Human Rights (ECHR),[226] Association of Southeast Asian Nations (ASEAN) Human Rights Declaration[227] and the Charter of Fundamental Rights of the European Union[228] are no more specific (aside from being grouped together with conscience and religious beliefs and their manifestation, as is the case also under the Convention on the Rights of the Child,[229] and the Declaration on the Elimination of All Forms of Intolerance and of Discrimination Based on Religion or Belief).[230] The American Convention on Human Rights (ACHR) is more specific, but

---

[222] John Tasioulas, *Saving Human Rights from Human Rights Law*, 52 VANDERBILT J. OF TRANSNAT'L L. 1167, 1193 (2019).

[223] G.A. Res. 217 (III) A, Universal Declaration of Human Rights, at 5 (Dec. 10, 1948).

[224] See, for example, the declarations and reservations of the US, notably: 'That the United States declares that the provisions of articles 1 through 27 of the Covenant are not self-executing.' International Covenant on Civil and Political Rights, Dec. 16, 1966, 999 U.N.T.S. 171.

[225] G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights art. 18 (Dec. 16, 1966).

[226] Convention for the Protection of Human Rights and Fundamental Freedoms art. 9, Nov. 4, 1950, 213 U.N.T.S. 221.

[227] ASEAN Human Rights Declaration gen. princ. 22, Nov. 19, 2012.

[228] Charter of Fundamental Rights of the European Union art. 10, Dec. 7, 2000, 2000/C 364/01.

[229] Convention on the Rights of the Child art. 14, Nov. 20, 1989, 1577 U.N.T.S 3.

[230] G. A. Res. 36/55, Declaration on the Elimination of All Forms of Intolerance and of Discrimination Based on Religion or Belief, at 1-3 (Nov. 25, 1981).

in its text does not separate freedom of thought from its manifestation in the form of discernible communication, linking it to free expression.[231] The African Charter on Human and Peoples' Rights (ACHPR) does not contain a provision on freedom of thought, but has one addressing freedom of conscience.[232]

Interpreters of these instruments are therefore central to understanding what the right to freedom of thought entails as a matter of law. But what do these actors say about this right?

1. International Machinery and Related Commentary

One starting point is examining the work of the UN Human Rights Committee. General Comment 22 states that the right to freedom of thought "does not permit any limitations whatsoever."[233] This understanding distinguishes freedom of thought from its manifestation, which engages other human rights that do permit limitations and restrictions being placed on them. The substance of this right also ties to Article 19(1) of the ICCPR, which concerns the right to hold opinions without interference.[234]

As the discussion regarding freedom of expression revealed, fact-check alerts and alternative source digital nudges do not appear to be in tension with imparting and receiving information. Instead, they have the potential to promote this aspect of the right by making information visible to social media users that would likely otherwise remain unseen on these platforms due to their algorithmic curation practices on newsfeeds. Digital nudges aimed at managing misinformation do have the potential to influence opinions; indeed, this is part of their purpose. However, they do not appear to interfere with the holding of opinions. Yet General Comment 22 suggests that the right to freedom of thought is inviolable, demanding absolute protection.[235] Opinions around the same time in the early 1990s and leading up to that period appear to align with this understanding,[236] encapsulated in the remark that this right is "the basis and the origin of all other [human] rights."[237] This makes sense, considering that people cannot

[231] American Convention on Human Rights art. 13, Nov. 22, 1969, 1144 U.N.T.S 123.

[232] African Charter on Human and Peoples' Rights art. 8, Jun. 27, 1981, 1520 U.N.T.S 217.

[233] H.R.C. *CCPR General Comment No. 22: Article 18 (Freedom of Thought, Conscience or Religion)*, ¶ 3, CCPR/C/21/Rev.1/Add.4 (Jul. 30, 1993).

[234] G.A. Res. 2200A (XXI), *supra* note 22, at art. 19(1).

[235] *See also id.* at art 4(2).

[236] Martin Scheinin, *Article 18*, *in* THE UNIVERSAL DECLARATION OF HUMAN RIGHTS: A COMMENTARY 263, 266 (Asbjørn Eide, et al. eds., 1992).

[237] U.N. Commission on Human Rights, 3rd Sess., 60th mtg. at 10, E/CN.4/SR.60 (Jun. 23, 1948).

freely do things, refrain from doing things, or guide such conduct with belief and conscience without first being able to think freely.

More recent insights come from Ahmed Shaheed in an interim report from 2021 on freedom of thought,[238] which finds that respecting the right includes "ensuring autonomy to develop thoughts, free from impermissible influences."[239] A question that therefore requires answering in the law is whether digital nudging, specifically fact-check alerts and alternative sources, are impermissible influences. Yet assessing this requires knowing what the constituent elements of the right to freedom of thought are, and how digital nudges measure-up against them. But as Shaheed notes, "little is clear about the right's core elements or 'attributes.'"[240] At present, four elements have been proposed:

> (a) not being forced to reveal one's thoughts; (b) no punishment and/or sanctions for one's thoughts; (c) no impermissible alteration of one's thoughts; and (d) States fostering an enabling environment for freedom of thought.[241]

The first three of these elements are altered versions of those already set out by Susie Alegre in 2017 drawing on the work of Ben Vermeulen:[242] "the right not to reveal one's thoughts or opinions; the right not to have one's thoughts or opinions manipulated; and the right not to be penalised for one's thoughts."[243]

The interfacing of digital nudges with social media users appears not to concern elements (a) and (b), and it is debatable whether the positive *legal* obligation associated with this right extends as far as that depicted in (d), as there seems to be a conflation here between human rights law and human interests and values.[244] The following analysis thus focuses on element (c). With respect to this element—that there be "no impermissible alteration of one's thoughts"—there are many legally permissible alterations to thought, such as educating people to eat healthy diets,

---

[238] Ahmed Shaheed (Special Rapporteur on Freedom of Religion or Belief), *Freedom of Thought*, U.N. Doc. A/76/380 (Oct. 5, 2021).

[239] *Id*. at 6.

[240] *Id*. ¶25.

[241] *Id*.

[242] *See* B. Vermeulen, *Freedom of Thought, Conscience and Religion*, *in* THEORY AND PRACTICE OF THE EUROPEAN CONVENTION ON HUMAN RIGHTS 752 (Pieter van Dijk, et al. eds., 2006).

[243] Susie Alegre, *Rethinking Freedom of Thought for the 21st Century*, 3 EUR. HUM. RTS. L. REV. 221, 225 (2017).

[244] *See* John Tasioulas, *Saving Human Rights from Human Rights Law*, 52 VAND. J. TRANSNAT'L L. 1167, 1179-95 (2019); *see also* HURST HANNUM, RESCUING HUMAN RIGHTS: A RADICALLY MODERATE APPROACH 119-134, 157-172 (2019).

indoctrinating kids in school,[245] advertising, and maybe, according to the former U.N. Special Rapporteur on freedom of religion or belief, "'nudges' to influence citizens' behaviour towards desired outcomes."[246]

Manfred Nowak noted the difficulty in distinguishing between legally permissible and impermissible influences on human thought, and that violations may be limited to situations where opinions are involuntarily influenced.[247] Digital nudging poses problems in this regard. There are currently three articulated categories of impermissible alterations of human thought within element (c), which — if present — violate the right to freedom of thought: coercion, modification of thought, and manipulation of thought.[248] Fact-check alerts and alternative source digital nudges do not meet the threshold of coercion, which is when an actor is effectively obliged to carry out conduct in a prescribed manner due to a threat being issued for non-compliance, or at least a negative incentive existing that limits the freedom to decide between options, thus compelling the undertaking of conduct in a particular way and removing any genuine choice to do things differently.[249] In the words of Joseph Raz:

> Coercion diminishes a person's options. It is sometimes supposed that that provides a full explanation of why it invades autonomy. It reduces the coerced person's options below adequacy. But it need not. One may be coerced not to pursue one option while being left with plenty of others to choose from.[250]

---

[245] Deana Heath, *British Empire is Still Being Whitewashed by the School Curriculum – Historian on Why This Must Change*, THE CONVERSATION (Nov. 2, 2018), https://theconversation.com/british-empire-is-still-being-whitewashed-by-the-school-curriculum-historian-on-why-this-must-change-105250.

[246] U.N. Secretary-General, *Freedom of Religion or Belief*, ¶ 28, U.N. Doc. A/76/380 (Oct. 5, 2021).

[247] MANFRED NOWAK, U.N. COVENANT ON CIVIL AND POLITICAL RIGHTS: CCPR COMMENTARY 442 (2005); Kate Jones, *Online Disinformation and Political Discourse: Applying a Human Rights Framework*, CHATHAM HOUSE INTERNATIONAL LAW PROGRAMME 1, 32-37 (2019), https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf.

[248] *See* Ahmed Shaheed (Special Rapporteur on freedom of religion or belief), *Freedom of religion or belief*, ¶ 29-39, U.N. Doc. A/76/380 (Oct. 5, 2021).

[249] Int'l Law Comm'n, Rep. on the Work of Its Twenty-Seventh Session, U.N. Doc. A/10010/Rev.1 at 69 (1975); Int'l Law Comm'n, Rep. on the Work of Its Thirty-First Session, U.N. Doc. A/34/10 at 93 (1979); Antonios Tzanakopoulos, *The Right to be Free from Economic Coercion*, 4 CAMBRIDGE INT'L L. J. 616, 618–623 (2015); James D. Fry, *Coercion, Causation, and the Fictional Elements of Indirect State Responsibility*, 40 VANDERBILT J. TRANSNAT'L L. 611 (2007); Denis G. Arnold, *Coercion and Moral Responsibility*, 38 AM. PHIL. Q. 53 (2001).; David A. Lake, *Authority, Coercion, and Power in International Relations*, *in* BACK TO BASICS: STATE POWER IN A CONTEMPORARY WORLD 55 (Martha Finnemore & Judith Goldstein eds., 2013); Bernd Beber & Christopher Blattman, *The Logic of Child Soldiering and Coercion*, 67 INT'L ORG. 65 (2013).

[250] Joseph Raz, THE MORALITY OF FREEDOM 377 (1988).

Raz also helps distinguish coercion from manipulation, as the latter, unlike the former, may "not interfere with a person's options. Instead, it perverts the way that person reaches decisions, forms preferences or adopts goals."[251] What a coercing actor requires can be predetermined to an extent, highlighting some commonality with the outcomes that can be predicted with varying degrees of accuracy when implementing nudges. However, there is a difference between undertaking conduct in settings where there exists a possibility, however strong or weak, that a punishment will ensue should the "wrong" option be chosen, versus settings where no such indication of possible punishment exists.

The next issue of thought modification also appears to not concern the use of the two digital nudges. This is because, from the current perspective of human rights law, thought modification is limited to the "direct alteration of brain chemistry or brain function" where the measure in question "bypasses psychological processes to directly alter biological function."[252] Even if one considered digital nudges to be a means of directly altering biological and chemical brain functions, to what extent states should protect against such processes remains unclear[253]—lest having a few too many pints at the pub becomes an unlawful practice because boozing can cause altered thought.[254] So long as there is the free, prior, and informed consent of individuals, perhaps even invasive procedures undertaken for the purpose of altering brain activity may not be contrary to this element of the right.[255]

Consent is also a key factor relating to manipulation of thought, the third currently articulated category of impermissible alterations of human thought. Whether such manipulation occurs as a matter of law is informed by assessing consent. As the discussion above set out, nudging in general engages psychological processes—considered by some to be a manipulative measure, or at least potentially manipulative if the nudge in

---

[251] *Id*. at 377–78.

[252] Shaheed, *supra* note 239, at ¶ 32-35; *see also* Agnieszka K. Adamczyk & Przemsław Zawadzki, *The Memory-Modifying Potential of Optogenetics and the Need for Neuroethics*, NANOETHICS 207, 207-08 (2020).

[253] *See generally* Andrea Lavazza, *Freedom of Thought and Mental Integrity: The Moral Requirements for Any Neural Prosthesis*, 12 FRONTIERS NEUROSCIENCE 1, 5 (2018).

[254] Beverly Merz, *This is your brain on alcohol*, HARV. HEALTH PUBL'G (Jul. 14, 2017), https://www.health.harvard.edu/blog/this-is-your-brain-on-alcohol-2017071412000.

[255] Shaheed, *supra* note 239, at ¶32-35; Agnieszka K. Adamczyk & Przemyslaw Zawadzki, *The Memory-Modifying Potential of Optogenetics and the Need for Neuroethics*, 14 NANOETHICS 207 (2020); Jan Christoph Bublitz, *Freedom of Thought in the Age of Neuroscience: A Plea and a Proposal for the Renaissance of a Forgotten Fundamental Right*, 100 ARCHIVES FOR PHIL. OF L. AND SOC. PHIL. 1, 9-11 (2014).

question is not designed with care.[256] Alegre points out that "[i]nfluence against the person's will or without their consent and methods that try to bypass a person's rational faculties to influence them are likely to be manipulative. In the digital world, the tools of technological influence are increasingly shaping our minds without us realising it."[257]

This account speaks to the purpose of digital nudging. Recall this measure is intended to interface with System 1 thinking, which does not appear to be rational, meaning System 2—the comparatively more rational type of thinking—is not exactly bypassed but may not be engaged in related decisions either. Drawing on the discussion above concerning whether decisions based on System 1 thinking are really choices, should a digital nudge not prompt System 2 thinking, however briefly, then it may be argued that the particular social media user exposed to that digital nudge was influenced without their awareness to make a decision that they might not have otherwise made. One example of such a nudge in the offline world is a cafeteria designed using choice architecture to guide consumers toward picking certain foods without them necessarily being aware that they are being influenced by where foods are placed. For example, ensuring a certain food is presented to a consumer before other options makes it more likely to be selected.[258]

On the one hand, should a digital nudge do something similar by leading a social media user to a particular course of conduct without any System 2 thinking involved, then freedom of thought may have been impermissibly influenced because the outcome occurred without the user undertaking any active part in the cognitive process, but instead passively doing so without conscious thought. On the other hand, as shown with respect to the alternative source digital nudge, its use appears to be geared toward prompting System 2 thinking. This measure would thus be the opposite of bypassing rational decision-making processes, and instead would promote such thinking that consciously takes into consideration other sources of information before making decisions regarding what to read and share. The related choice would arguably be more likely to be free

---

[256] *See* Christopher McCrudden & Jeff King, *The Dark Side of Nudging: The Ethics, Political Economy, and Law of Libertarian Paternalism*, in CHOICE ARCHITECTURE IN DEMOCRACIES: EXPLORING THE LEGITIMACY OF NUDGING 67, 75 (Alexandra Kemmerer et al. eds., 2017).

[257] SUSIE ALEGRE, FREEDOM TO THINK: THE LONG STRUGGLE TO LIBERATE OUR MINDS 28 (2022).

[258] Elizabeth Velema, et al., *Using Nudging and Social Marketing Techniques to Create Healthy Worksite Cafeterias in the Netherlands: Intervention Development and Study Design*, BMC PUB. HEALTH 1, 2 (2017).

because the person is aware of the thought process involved, in what can be described as cognitive consciousness.[259]

It is worth noting here that this process of thinking may be why digital nudges can only be effective at reducing the spread of misinformation but perhaps not disinformation. Should users intend to create and/or spread false information, a digital nudge may well be of no use in terms of prompting consideration of a different choice,[260] much in a similar way that if a consumer intends to have a burger for lunch no amount of deliberately constructed choice architecture is going to nudge them into eating a salad when they arrive at the cafeteria.

Karen Yeung has set out some concerns regarding the combined analytic prowess of Big Data with the use of nudges, underscoring the pervasiveness of digital nudging on social media platforms as a tool that if improperly handled could be used for wrongful purposes, with troubling implications for democracies.[261] Research suggests that thought manipulation is that which interferes with how human understanding is shaped in order to bring about the formation of thoughts that align with particular worldviews.[262] Is this not a root rationale behind nudging attempting to shape conduct that is in the interests of a dominant group, which may or may not align with those of a dominated group?[263] Power asymmetry is a component of whether the potential to manipulate thought exists.[264] Such asymmetry is present between those implementing digital nudges on newsfeeds and the users subjected to them.[265] In such unbalanced relations, there is the potential to manipulate thought, especially considering "the data collected about users will reflect things that they may not even be consciously aware of themselves."[266]

That said, should the intended influence of fact-check alerts and alternative sources be clearly articulated by the social media platforms

---

[259] Robert Van Gulick, *Consciousness and Cognition*, *in* THE OXFORD HANDBOOK OF PHIL. OF COGNITIVE SCI. 19 (Eric Margolis, Richard Samuels. and Stephen P. Stich eds., 2012).

[260] *See* Mathias Osmundsen, et al., *Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter*, 115 AM. POL. SCI. REV. 999 (2021).

[261] *See* Karen Yeung, *"Hypernudge": Big Data as a Mode of Regulation by Design,* 20 INFO., COMMC'N & SOC'Y 118 (2017).

[262] Teun A. Van Dijk, *Discourse and Manipulation*, 17 DISCOURSE & SOC'Y 359 (2006).

[263] *Id*.

[264] Shaheed, *supra* note 239, at ¶ 36.

[265] Dipayan Ghosh & Nick Couldry, *Digital Realignment: Rebalancing Platform Economies from Corporation to Consumer,* M-RCBG Associate Working Paper No. 155 (2020), https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp155.

[266] Cameran Ashraf, *Exploring the Impacts of Artificial Intelligence on Freedom of Religion or Belief Online,* 26 INT'L J. OF HUM. RTS. 757, 770-4 (2022).

implementing them, meaning such transparency could fulfill at least the tacit consent of users being subjected to these digital nudges, the "concealment or obfuscation" that arguably contributes to a measure being incompatible with the right to freedom of thought becomes less likely.[267] Therefore, in order for social media companies to not risk allegations that the use of digital nudges are contrary to the right to freedom of thought, they could explicitly state what any digital nudges are being used for and why, such as attempting to reduce the spread of misinformation and explaining the basis behind an approach utilizing these behavioral interventions. The informed consent of users is a factor to consider in assessments regarding not only freedom of thought, but those including freedom of expression and privacy as well.[268]

Such notifications to users, however, should not be hidden away in "incomprehensible" terms and conditions of service or privacy policies,[269] but instead presented in plain language within the respective platform where ease of access is not an issue. It is time to move on from burdening social media users with a "duty to read the unreadable."[270] This accessibility and readability is important considering how content appears to social media users on their newsfeeds because of the current workings of algorithmic information curation. If designed and implemented with care, digital nudges have the potential to assist in protecting the right to freedom of thought by changing and expanding the information presented to users, instead of being a measure that is contrary to this human right. As noted in the 2018 report of the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression: "To be sure, all sorts of social and cultural settings may limit an individual's exposure to information. But by optimizing for engagement and virality at scale, AI-

[267] Ahmed Shaheed (Special Rapporteur), *Freedom of religion or belief*, ¶ 36, U.N. Doc. A/76/380 (Oct. 5, 2021).

[268] *See* TOBY MENDEL, ET AL., GLOBAL SURVEY ON INTERNET PRIVACY AND FREEDOM OF EXPRESSION 23, 44, 97-101 (2012).

[269] Kevin Litman-Navarro, *We Read 150 Privacy Policies. They Were an Incomprehensible Disaster*, N.Y. TIMES (June 12, 2019), https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html; Dustin Patar, *Most Online "Terms of Service" Are Incomprehensible to Adults, Study Finds*, VICE (Feb. 12, 2019), https://www.vice.com/en/article/xwbg7j/online-contract-terms-of-service-are-incomprehensible-to-adults-study-finds.

[270] Uri Benoliel & Shmuel I. Becher, *The Duty to Read the Unreadable*, 60 B.B. L. REV. 2255 (2019).

assisted personalization may undermine an individual's choice to find certain kinds of content."[271]

As the discussion above revealed, the alternative source nudge has the potential to counteract this problem. Instead of negatively affecting intellectual freedom and critical thinking by minimizing exposure to diverse views, this measure can form part of the opposite practice, whereby people are exposed to new sources that may assist in less credence being lent to misinformation.[272] By exposing users to different viewpoints than those they might otherwise encounter while on social media platforms, this particular digital nudge can arguably promote the right to freedom of thought while people expose themselves to this part of the digital realm.

This is a key difference between fact-check alerts and alternative source digital nudges in terms of their use being contrary to, or compatible with, the right.[273] The former prompts users to consider whether the content they are accessing is factually accurate, whereas the latter encourages users to read from different sources while remaining agnostic with respect to their content and whether it is accurate by not making any explicit statement on the matter. The difference between these two digital nudges may therefore impact freedom of thought differently and, consequently, whether a judgment considers them to be compatible with the right when considered separately or collectively. Variance in decisions on their legality may well stem from the premise that fact-check alerts influence the manner in which users understand information presented to them, rather than what they choose to read out of a selection of options and then understand afterwards without further input. Such is the case with alternative sources of information being presented through digital nudges without an explicit indication being provided as to what source is factually accurate or not.

An additional component in this process of using alternative source digital nudges could be adding another review prompt. Once the platform has recognized that the user has clicked through onto at least one of the alternative sources presented to them, should that user still decide to share the original content containing misinformation, a further interstitial webpage could pop-up asking them if they are sure they would like to share that source. Such a nudge would be more clearly questioning the accuracy

[271] UNGA, Promotion and Protection of the Right to Freedom of Opinion and Expression, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, ¶ 12, A/73/348 (Aug. 29, 2018).

[272] Ahmed Shaheed (Special Rapporteur), *Freedom of Thought*, Interim Report of the Special Rapporteur on Freedom of Religion or Belief, para. 67 – *see also* ¶ 73-75, A/76/380 (Oct. 5, 2021).

[273] With many thanks to Kate O'Regan and Katie Pentney for helping me understand this point more clearly.

of the original source without necessarily explicitly indicating that it contains falsehoods. A key balancing act for social media platforms would be to ensure that any such double-tap digital nudges effectively reduce the sharing of misinformation without reaching the threshold of being so recurrent on newsfeeds that users become frustrated; possibly frustrated enough to decrease their engagement on the platform, or leave it, whether permanently or temporarily, including potentially for another platform that may be a competitor. Such abandonment may also do little if anything to stem the flow of misinformation and could even result in its increase if there was an exodus of users from one platform, because alternative platforms may allow users that switched more space from content curation and moderation mechanisms that annoy them.[274]

2. Perspectives from Case Law

In addition to these insights gained from the work of UN mechanisms and surrounding research, related case law requires investigation. Case law at the international level under the ICCPR framework is currently not in a position to assist in ascertaining the contours of the freedom of thought any further than those already articulated. There appears to be only one case that has referred to it.[275] The UN Human Rights Committee has also yet to provide meaningful engagement with the right to freedom of thought when complaints have been submitted regarding alleged violations, instead dismissing these claims in lieu of findings confirming violations of the rights to freedom of association and expression.[276] Such decisions send a message that this right is of comparatively lesser importance than the others associated with its manifestation, which seemingly runs contrary to the right being apparently absolute insofar as unmanifested thought is concerned.

A similar approach appears to have been taken in matters lodged under regional human rights instruments. In terms of regional case law, there have been few developments. With respect to matters under the ACHR, Cláudio de Oliveira Santos Colnago and Bethany Shiner find that the right to freedom of thought "is not explicitly recognised as having any value beyond its role in fueling expression," meaning related case law "has made no attempt to delineate the inner realm of thought from external

[274] With thanks to Taylor Desgrosseilliers for raising this point.

[275] UNHRC, Communication, ¶ 3.2, 7.2, No. 878/1999, CCPR/C/78/D/878/1999 (Jul. 16, 2003).

[276] UNHRC, Communication, ¶ 7.4, No. 1119/2002, CCPR/C/84/D/1119/2002 (Aug. 23, 2005); UNHRC, Communication, ¶ 10.5, No. 628/1995, CCPR/C/64/D/628/1995 (Nov. 3, 1998).

manifestations."[277] Where there is case law under the ACHR that indirectly concerns freedom of thought, little of it reveals any details that could pose challenges regarding the compatibility of this right with the use of digital nudges beyond those that have already been analyzed.[278]

However, at least two decisions are noteworthy. In one, the Inter-American Court of Human Rights ruled on the principle of human autonomy, which, according to the Court, "prohibits any State action that attempts to 'instrumentalize' individuals," or, "in other words, convert them into a means for purposes unrelated to their choices about their own life, body and full development of their personality."[279] In the other case, a reading of what constitutes privacy and the state's legal obligation to protect against inferences with it by non-state actors holds that "[t]he scope of privacy is characterized as being free and immune to invasions or abusive or arbitrary attacks by third parties or public authority and may include, among other dimensions, the freedom to make decisions related to various areas of a person's life."[280]

Reflecting on these combined rulings, there is a perceptible tension with digital nudges in the form of both fact-check alerts and alternative sources. However, whether this tension with respect to privacy amounts to these measures being unlawful under the ACHR is doubtful, considering inferences must be "arbitrary or abusive" in order to reach this threshold.[281] There are many unknowns about whether digital nudges are contrary to the right to freedom of thought under the ACHR, which calls for it to "be

---

[277] Cladio de Oliveira Santos Colnago & Bethany Shiner, *A Distinct Right to Freedom of Thought in South America: The Jurisprudence of the Inter-American Court of Human Rights, Neurotechnology and the Application of Bioethics Principles*, 8 EUR. J. OF COMPAR. L. AND GOVERNANCE 245, 246 (2021).

[278] Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism, Advisory Opinion OC-5/85, Inter-Am. Ct. H.R. (ser. A) No. 5 (Nov. 13, 1985); Ivcher-Bronstein v. Peru, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 74 (Feb. 6, 2001); Olmedo Bustos v. Chile, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 73 (Sept. 5, 2001); Herrera-Ulloa v. Costa Rica, Preliminary Objections, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 107 (July 2, 2004); Canese v. Paraguay, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 111 (Aug. 31, 2004); Palamara Iribarne v. Chile, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 135 (Nov. 22, 2005); Claude-Reyes v. Chile, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 151 (Sept. 19, 2006); Kimel v. Argentina, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 177 (May 2, 2008).

[280] I.V. v. Bolivia, Preliminary Objections, Merits, Reparations and Costs, Inter-Am. Ct. H.R. (ser. C) No. 329, ¶ 150 (Nov. 30, 2016).

[280] Fontevecchia v. Argentina, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 238, ¶ 48 (Nov. 29, 2011).

[281] Organization of American States, American Convention on Human Rights, Nov. 22, 1969, Art. 11(2).

interpreted as a separate and distinct right" so that it becomes possible "to specify the exact contours that the right may have."[282]

Not much is different under the ECHR. But again, there are insights to be gained from some outputs, and not only those from the European Court of Human Rights. A declaration by the Committee of Ministers of the Council of Europe states that "fine grained, sub-conscious and personalised levels of algorithmic persuasion may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions. These effects remain underexplored but cannot be underestimated."[283]

The Committee of Ministers goes on to draw attention to "the right of human beings to form opinions and take decisions independently of automated systems," and the threats posed by the "capacity to use personal and non-personal data to sort and micro-target people, to identify individual vulnerabilities and exploit accurate predictive knowledge, and to reconfigure social environments in order to meet specific goals and vested interests."[284] In addition, initiatives that have the potential to treat the root causes of misinformation and its spread have been encouraged, including "empowering users by promoting critical digital literacy skills and robustly enhancing public awareness" about data generation, collection and use.[285]

Although these insights are not a reflection of positive law, they are an indicator of the political winds that the law might follow under the ECHR framework. Upon reflecting on these concerns, the use of digital nudges on social media newsfeeds raises flags. Nudges can influence opinions and are tailored to meet predetermined outcomes. But whether the forms of fact-check alerts and alternative sources do so to the extent that they are unlawful under the ECHR framework depends on what the Court has to say on freedom of thought.

Despite many aspects of the digital realm raising "urgent questions not just about privacy but also the protection of our thoughts and ability to think,"[286] ECHR case law has yet to address this matter in any detail. There is little elaboration as to what measures would be incompatible with this right, even if there are statements indicating its absolute nature under the

---

[282] Claudio de Oliveria Santos Colnago & Bethany Shiner, *A Distinct Right to Freedom of Thought in South America*, 8 EUR. J. OF COMPAR. L. AND GOVERNANCE 245, 270 (2021).

[283] Eur. Consult. Ass., *Declaration by the Comm. of Ministers on the Manipulative Capabilities of Algorithmic Processes*, 1337th Meeting of the Ministers' Deputies, ¶ 9 (2019).

[284] *Id*.

[285] *Id*. at sub-para. (e).

[286] Patrick O'Callaghan & Bethany Shiner, *The Right to Freedom of Thought in the European Convention on Human Rights*, 8 EUR. J. COMPAR. L. GOVERNANCE 112, 120 (2021).

ECHR.[287] Yet it is not easy to understand what is absolute about the right under this legal framework regarding unmanifested thoughts because "the right to freedom of unmanifested thought, outside of the freedom of religious or philosophical belief contexts, has not yet been argued before the ECtHR."[288] That said, as is also the case under the ICCPR framework, coercion attempting to alter thought is prohibited—meaning if digital nudges are proven to be a coercive measure then their use would be incompatible with the law under the ECHR.[289] Failure to protect against any such coercion by non-state actors would also likely put a state bound by this treaty in breach of its positive legal obligation.[290]

A further ground of possible contention is that any measure that can be construed as psychological "deprogramming" is incompatible with the right.[291] Considering the progressive deradicalization problem, it could be that prolonged exposure to digital nudges on newsfeeds falls within the scope of deprogramming. As such, it may be that the implementation of digital nudges on social media newsfeeds is best strictly limited to periods of time where there are influxes of misinformation on the platform in question. Their use would thus be phased out along with dropping numbers of sources containing misinformation, only to be re-introduced should there be further spikes in misinformation circulation.

3. Situating the U.S. Approach and Future Constitutional Challenges

The right to freedom of thought remains a collage across domestic legal systems as well. Previous research shows that in a randomly selected sample of states across continents, freedom of thought is embedded in many constitutions.[292] The U.S. is of particular interest because the

[287] Kosteski v. The Former Yugoslav Republic of Macedonia, App. No. 55170/00, ¶ 39 (Apr. 13, 2006), https://hudoc.echr.coe.int/eng?i=001-73342; Kokkinakis v. Greece, App. No. 14307/88, ¶ 33 (May 25, 1993), https://hudoc.echr.coe.int/eng?i=001-57827; C. v. The United Kingdom, App. No. 10358/83, p. 147 (Dec. 15, 1983); Alexandridis v. Greece, App. No. 19516/06, ¶ 38 (Feb. 21, 2008), https://hudoc.echr.coe.int/eng?i=001-85188; Grzelak v. Poland, App. No. 7710/02, ¶ 87 (June 15, 2010), https://hudoc.echr.coe.int/eng?i=001-99384; Sinan Işik v. Turkey, App. No. 21924/05 (Feb. 2, 2010), https://hudoc.echr.coe.int/eng?i=001-97087.

[288] O'Callaghan & Shiner, *supra* note 287, at 131-132.

[289] *See* Ivanova v. Bulgaria, App. No. 52435/99, ¶ 79 (Apr. 12, 2007) https://hudoc.echr.coe.int/fre?i=001-80075; Mockutė v. Lithuania, App. No. 66490/09, ¶ 119, 129 (Feb. 27, 2018), https://hudoc.echr.coe.int/eng?i=001-181202.

[290] Eweida and Others v. United Kingdom, App. No. 36516/10, ¶ 84 (Jan. 15, 2013), https://hudoc.echr.coe.int/eng?i=001-115881.

[291] Riera Blume and Others v. Spain, App. No. 37680/97 (Oct. 14, 1999), https://hudoc.echr.coe.int/fre?i=002-6630.

[292] Davit Harutyunyan & Lilit Yeremyan, *Freedom of Thought: Legal Protection from Manipulation*, 14 WISDOM 131, 135-137 (2020).

companies providing a number of (currently) popular social media platforms reside there.[293] The U.S. Supreme Court views the right as the starting point for having and exercising freedom.[294] In *Jones v. Opelika*, the Court held:

> Freedom of speech, freedom of the press, and freedom of religion all have a double aspect—freedom of thought and freedom of action. Freedom to think is absolute of its own nature; the most tyrannical government is powerless to control the inward workings of the mind. But even an aggressive mind is of no missionary value unless there is freedom of action—freedom to communicate its message to others by speech and writing. Since, in any form of action, there is a possibility of collision with the rights of others, there can be no doubt that this freedom to act is not absolute, but qualified, being subject to regulation in the public interest which does not unduly infringe the right.[295]

In *Palko v. Connecticut*, freedom of thought was linked to free speech as a "matrix, the indispensable condition, of nearly every other form of freedom."[296] "Liberty of the mind" was considered to warrant protection as much as the "liberty of action."[297]

The link between these rights under U.S. law is critical because freedom of thought is not directly protected.[298] Yet the indirect protection that this right receives is limited to the marketplace of ideas approach supported by the Supreme Court. According to the marketplace of ideas theory, manifested thoughts should be left to compete with each other without interference from the state unless such expression poses a "clear and present danger."[299] This position also means actors considered to be in the private sphere can lawfully moderate expression, which is why measures such as content removal and deplatforming can be considered legally permissible. The implementation of such measures by social media

---

[293] *See* Cristina Criddle, *Instagram Cuts Influencer Payments for Short Videos*, FIN. TIMES (April 5, 2022), https://www.ft.com/content/df2d753d-4346-49d8-88d7-441156dc8dc3.

[294] Ashcroft v. Free Speech Coalition, 535 U.S. 234, 253 (2002) ("The right to think is the beginning of freedom").

[295] Jones v. Opelika, 316 U.S. 584, 618 (1942).

[296] Palko v. Connecticut, 302 U.S. 319, 327 (1937).

[297] *Id*.

[298] *See generally* John G. Francis & Leslie Francis, *Freedom of Thought in the United States: The First Amendment, Marketplaces of Ideas, and the Internet*, 8 EURO. J. OF COMPAR. L. AND GOVERNANCE 192 (2021).

[299] Schenck v. United States, 249 U.S. 47 (1919); Debs v. United States, 249 U.S. 211 (1919); Frohwerk v. United States, 249 U.S. 204 (1919); Schaefer v. United States, 251 U.S. 466 (1920); Pierce v. United States, 252 U.S. 239 (1920); Whitney v. California, 274 U.S. 357 (1927); Dennis v. United States, 341 U.S. 494 (1951).

platforms is considered an exercise of the companies' rights to express their views.[300]

Yet there are those that consider certain social media platforms to be the "functional equivalent of a public forum" and are concerned about content removal and deplatforming being arbitrary.[301] Digital nudging is a less blunt way of balancing owning companies' rights to free expression under current U.S. law with those of users, while factoring in the responsibilities these companies have to protect against their products and services leading to harm, to which misinformation can contribute.

There appears to be little U.S. law that would render the use of fact-check alerts and alternative sources by social media platforms contrary to the right to freedom of thought. A possible exception is the position taken in the *Citizens United* case, where the Court stated:

> Speech restrictions based on the identity of the speaker are all too often simply a means to control content.
>
> Quite apart from the purpose or effect of regulating content, moreover, the Government may commit a constitutional wrong when by law it identifies certain preferred speakers. By taking the right to speak from some and giving it to others, the Government deprives the disadvantaged person or class of the right to use speech to strive to establish worth, standing, and respect for the speaker's voice. The Government may not by these means deprive the public of the right and privilege to determine for itself what speech and speakers are worthy of consideration.[302]

The significance of this judgment is that both the digital nudges examined above extend preference to content that is not considered to be misinformation, and are measures intended to influence social media users' decisions regarding what sources are worthy of their consideration. John Francis and Leslie Francis mention, "regulations that may be permissible because they are content-neutral may be struck down because of their broader effects on speech by commercial actors, as with *Citizens United*."[303] Although this case indirectly concerns freedom of thought, should the U.S. government—whether through the Federal Trade Commission in

---

[300] Thomas A. Berry & Nicole Saad Bembridge, *The First Amendment Protects Everyone, Even Facebook and Twitter*, CATO INST., (Nov. 22, 2021), https://www.cato.org/commentary/first-amendment-protects-everyone-even-facebook-twitter.

[301] Elijah O'Kelley, *State Constitutions as a Check on the New Governors: Using State Free Speech Clauses to Protect Social Media Users from Arbitrary Political Censorship by Social Media Platforms*, 69 EMORY L. J. 111 (2019).

[302] Citizens United v. Federal Election Commission, 558 U.S. 310, 341 (2010).

[303] Francis, *supra* note 299, at 225.

accordance with the Social Media Nudge Act (if enacted) or otherwise—attempt to regulate social media newsfeeds by utilizing digital nudges, then there would exist grounds for a legal challenge.

However, such grounds for challenge do not currently extend to social media platforms implementing these measures themselves, considering the law according to the Supreme Court "does not prohibit *private* abridgement of speech" but "prohibits only *governmental* abridgment of speech."[304] A dissent in this case nevertheless highlights that the private operators of publicly accessible television channels carry out a public function established by the state.[305] The concurring opinion of Justice Clarence Thomas in a separate case concerning freedom of expression on social media is also relevant in this regard, where, in sum, he opined, "Respondents have a point . . . that some aspects of Mr. Trump's account resemble a constitutionally protected public forum. But it seems rather odd to say that something is a government forum when a private company has unrestricted authority to do away with it."[306] This consideration highlights some tension that appears to require further discussion, including whether social media platforms should be considered "common carriers." The related understanding here is that social media platforms could be treated similarly to transportation services available to the public, as they "carry" information across them from one user to another, meaning the state should be able to regulate them, including in order to prohibit content moderation by these platforms.[307] However, such a position may be at odds with the Communications Decency Act, which states, among other things, that "[n]o provider or user of an interactive computer service shall be held liable on account of . . . any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[308]

There appears to be a categorization problem in U.S. law regarding whether social media platforms are distributors of information, and arguably more similar to "common carriers" (such as buses, planes and trains); or publishers or speakers of information, and arguably more similar to newspapers and the like with editorial oversight and discretion; or a mix

[304] Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1928 (2019) (emphasis original).

[305] *Id.* at 1934 (Justice Sotomayor, dissenting).

[306] Biden v. Knight First Amend. Inst. (2021), 141 S. Ct. 1221, 1224 (Justice Thomas, concurring).

[307] *Id.* at 1224.

[308] 47 U.S.C. § 230.

of these categories.[309] Whatever categorization is ultimately settled upon will inform determinations on whether social media companies are protected under constitutional law to freely express themselves, such as by removing content that they do not want on their platforms, or whether the state will have authority to regulate these communication platforms, which cannot be said to be purely private or public. With the law here appearing to be in flux,[310] there could be a shift from the present position of private platforms on the internet not being considered public forums, toward a position where understandings about "traditional" public functions dissolve, which would fit with a trend, and not only in the U.S., where many "public functions" are now outsourced to non-state actors or privatized.

The distinction between the public and private spheres across states in today's world is no longer clear-cut; non-state actors now carry out numerous roles on the behalf of states, both officially and unofficially.[311] U.S. domestic law currently sides with social media platforms implementing digital nudges for the purposes of addressing misinformation, but may not permit the state to regulate these platforms in this way.

Whether at the domestic level or taken to regional judicial bodies or international quasi-judicial bodies, strategic litigation has the potential to help "clarify the scope, limit and application of the right [to freedom of thought]."[312] When examining the different legal approaches to this right, there is noticeable fragmentation. Whether from domestic legal machinery influencing the development of international law or vice versa, the right to freedom of thought requires further elaboration by bodies tasked with such endeavors. Considering the apparent absolute nature of the right under the ICCPR and ECHR frameworks, another difficult question arises about whether limitations on freedom of expression by extension limit freedom of

---

[309] *See Id.*; *See also* Jon Brodkin, *Texas Cites Clarence Thomas to Defend its Social Media Law*, WIRED (May 21, 2022, 8:00 AM), https://www.wired.com/story/texas-clarence-thomas-opinion-social-media-moderation-ban/.

[310] *See also* the seemingly conflicting decisions in Rumsfeld v. Forum for Acad. & Institutional Rts. 547 U.S. 47 (2006); and NetChoice v. Paxton 49 F.4th 439 (5th Cir. 2022).

[311] Some examples include for hire hackers, domestic bus and train companies, private military and security companies, and technology companies providing contact-tracing applications in addition to other products and services. *See* RICHARD MACKENZIE-GRAY SCOTT, STATE RESPONSIBILITY FOR NON-STATE ACTORS: PAST, PRESENT, AND PROSPECTS FOR THE FUTURE 38-69, 240-42 (2022).

[312] Bethany Shiner & Patrick O'Callaghan, *Introduction to a Comparative Study of the Right to Freedom of Thought*, 8 EUR. J. OF COMPAR. L. AND GOVERNANCE 107, 109 (2021); *See also* Susie Alegre, *Protecting Freedom of Thought in the Digital Age*, DIGITAL FREEDOM FUND (Mar. 23, 2020), https://digitalfreedomfund.org/protecting-freedom-of-thought-in-the-digital-age/.

thought. This calls into question the latter right's absoluteness, which is not recognized as such in domestic legal systems (such as Canada's).[313] This position fits with the text under the applicable provision of the ACHR and its accompanying case law, which sets out that the rights of freedom of thought and expression can be subject to impositions established by law that include protecting public order or health where necessary.[314] From the local to the international, there is a need for makers and shapers of the law at all levels of governance to clearly delineate the contours of this right from those rights with which it is associated. Not only will such efforts grant clarity surrounding the specific matter regarding the lawfulness of using digital nudges on social media newsfeeds to reduce the spread of misinformation, but they can also set out a path for other technological innovations to follow that ensures compliance with a right that goes to the very heart of surviving or thriving as an autonomous human being.

B. Surpassing What the Law Currently Says in Order to Help Develop It

The proclaimed importance of the right to freedom of thought does not sit well with the fact that a distinct provision within applicable legal instruments does not currently exist, even if it is considered to be a separate right from those concerning beliefs and conscience. While all three of these rights concern the internal workings of human minds, there is a lack of articulation in the law regarding freedom of thought, hence the calls "to further clarify the freedom's scope and content, including through a general comment."[315] This lack of clarity is arguably "undermining its practical application."[316] However, the current state of affairs also provides an opportunity for commentators to inform law-making bodies about how the law here should develop. There is a wealth of work to draw from that extends beyond the confines produced by doctrinal legal research. This work can, and arguably should, be taken forward by residents within states to push for changes that they consider appropriate. Such contestation and the related reasoning involved have the promise of weeding out the same ideas that can also constitute misinformation or contribute to it being lent credence.

---

[313] R. v. Sharpe, 1 SCR 45, 108 (2001); *See also* Dwight Newman, *Freedom of Thought in Canada: The History of a Forgetting and the Potential of a Remembering*, 8 EUR. J. OF COMPAR. L. AND GOVERNANCE 226, 226 (2021).

[314] American Convention on Human Rights art. 13(2), Nov. 22, 1969, 1144 U.N.T.S. 17955.

[315] Ahmed Shaheed (Special Rapporteur), *Interim report of the Special Rapporteur on freedom of religion or belief*, Summary, U.N. Doc. A/76/380 (Oct. 5, 2021).

[316] *Id.* at 14.

The answers held in the law at present indicate the need to develop the law alongside contemporary changes in societies, government, and industry. The right to freedom of thought across international, regional, and domestic law is somewhat clumsily lumped together with its manifestation. In an era where the innermost workings of the human mind are being extracted for profit, it is arguably past time to set out a clear and distinct understanding of the human right to freedom of thought. It has been asserted that social media platforms in particular are on a mission to get inside the heads of users through the extraction of ever-more bits of their personal data, making it imperative that the human right to freedom of thought is practically protected.[317] This includes when implementing measures that are intended to help address societal issues such as misinformation.

It was once said, "for every complex human problem, there is a solution that is clear, simple and wrong."[318] There is a need for nuance when considering measures that have the potential to benefit societies at scale (such as the use of digital nudging on social media newsfeeds in order to manage misinformation), but which simultaneously have the potential to undermine the common good instead of serve it. Human rights and the law enshrining them exist in part to facilitate this process. There need not be conflict between autonomy and community.[319] Jean-Jacques Rousseau was of the view that "[e]ach man, while detaching his own interests from the common interest, sees clearly that he cannot separate them entirely."[320] John Stuart Mill was also of the view that although an individual should not be compelled to undertake conduct that is against their will, there are good reasons for persuading them to undertake conduct that is right or better for them and their community in the pursuit of preventing harm.[321] Over two hundred years before these insights, Thomas Hobbes wrote that the freedom to think in order to form an "invisible" "internal faith" should be "exempted from all human jurisdiction."[322] Yet this process is no longer invisible to certain actors in the technology industry that run businesses

[317] Susie Alegre, *Protecting Freedom of Thought in the Digital Age*, CTR. FOR INT'L. GOVERNANCE INNOVATION (May 2021), https://www.cigionline.org/static/documents/PB_no.165.pdf.

[318] *South Africa Divestment: Hearing and Markups*, 98th Cong., 123 (1984) (statement quoting Henry Louis Mencken in H. Comm. D.C., Subcomm. Fiscal Aff. and Health).

[319] Duncan Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685, 1733-1735 (1976).

[320] JEAN-JACQUES ROUSSEAU, DISCOURSE ON POLICIAL ECONOMY AND THE SOCIAL CONTRACT 135 (Christopher Betts trans., 1994).

[321] JOHN STUART MILL, ON LIBERTY 22-23 (1863).

[322] THOMAS HOBBES, LEVIATHAN 354 (1651).

based on learning from personal data and exploiting the associated information in order to garner more. Although individual autonomy is currently alive across the globe, it is under threat. Keeping it safe depends on maintaining the ability of humans to form their own thoughts and make decisions based on those thoughts. This human agency is at tension with behavioral interventions that are intended to steer people in particular directions of conduct, even if, or perhaps especially because, there is little conscious thinking involved.

In order to ensure that managing misinformation on social media newsfeeds by utilizing digital nudges does not overstep in its interaction with the right to freedom of thought, the law needs to be articulated in clearer terms. It is all well and good to say "responses to the spread of disinformation and misinformation must be grounded in international human rights law,"[323] but if the law provides few answers about what measures are legally permissible interferences because they are, for example, legitimate, necessary, and proportionate in comparison to alternatives, then the law needs to mature before it can shoulder such a role. The law on the right to freedom of thought must grow out of its infancy, as it currently lacks the muscle to be maneuvered through an online world where the means to batter, bruise, and belittle it are robust.

IV. CONCLUSION: RECONCILABLE RIGHT NOW

There is something perverse about trying to tinker with what goes on in the human mind for the purpose of attempting to guide decision-making and influence conduct toward a predetermined outcome. Yet putting a finger on what precisely this is by reference to what the law has to say on the human right to freedom of thought is difficult at present. The right often being linked to other rights does not help with this uncertainty, and if the above analysis has shown anything, it is that clarity is needed. Further articulation and careful consideration of this right is called for at all levels of governance, from the local to the international.

It may be that the prohibition on psychological "deprogramming" under the ECHR framework could be read as prohibiting *prolonged* exposure to digital nudges, because such implementation risks guiding users to more readily accepting predetermined viewpoints over time. The narrowing of exposure to sources of information would be a related concern here, due to regression in terms of access to content and lack of

[324] Irene Khan (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), *Disinformation and Freedom of Opinion and Expression*, ¶ 30, U.N. Doc. A/47/25 (Apr. 13, 2021); *See also id*. at ¶ 31.

freethinking as a conceivable consequence. Mitigating such potential risks means limiting the use of this measure on social media to periods where there are influxes of misinformation on a particular platform, where digital nudges can be phased-out alongside reductions in misinformation, to be re-introduced contingent on there being further spikes.

In addition, should nudging be proven to be a coercive measure, then any nudges would be incompatible with the law under at least the ICCPR and ECHR frameworks. Within the U.S., should the state attempt to regulate misinformation through the implementation of digital nudges on newsfeeds, doing so may result in legal challenges and the related regulations being struck down in the courts, unless perhaps courts view misinformation on social media as a clear and present danger to society. However, the utilization of digital nudges by social media platforms is compatible with the current law in the U.S., whether this measure is used to manage misinformation or for other purposes connected to the right of companies to freely exercise their expressions.

It is not certain that the two digital nudges of fact-check alerts and alternative sources are contrary to human rights law as it stands. This measure for managing misinformation on social media platforms therefore appears to be legally permissible at present, which can be said with some confidence when referring to the international and regional legal frameworks of the ICCPR, ACHR, and ECHR, in addition to the domestic legal framework in the U.S. If designed with care, the alternative source digital nudge could actually promote freedom of thought. It could introduce social media users to sources of information that they would not normally encounter, at least online, thereby encouraging wider reading on subjects where misinformation is prevalent, and thus potentially helping dispel falsehoods gradually over time by exposure to new information.

However, there is acknowledgement that the right to freedom of thought is absolute under at least the ICCPR and ECHR frameworks, generating impressions that it cannot be interfered with at all. In principle, digital nudging could be incompatible with this inviolability understanding of the right. But this notion of the right's apparent absoluteness is at odds with these very same legal frameworks (and others) indicating that there are permissible influences and alterations to human thought that would not violate the right. Such contradictions are confusing, providing considerable grounds for contestation, and thereby benefiting powerholders by allowing them to implicitly stake a claim regarding what they consider to be lawful through their conduct and accompanying silence on matters of law unless or until they are challenged.

If the law is ever going to be in a position to practically protect the human right to freedom of thought, then it needs to develop. Whether digital nudges in the form of fact-check alerts and alternative sources remain arguably lawful measures for managing misinformation on social media newsfeeds is thus a different question. Laws change with the appetite for bringing such change about, whether through courts providing new and necessary insights into how the law should apply, or legislatures and governments looking to introduce new rules or complements to existing ones.

Yet whether the law should change is a matter to be hashed out within and across states. While courts may yet provide clarity as to the scope, substance, and applicability of the right to freedom of thought as it concerns digital nudging, whether new rules are needed and should be introduced is for residents of states to debate. If democracy is to be realized, this will mean that resulting views are taken onboard and represented by politicians and social media companies when advocating for the related interests at stake, in addition to their own.

There are complementary steps that can also be taken in order to advance these efforts. One is the Oversight Board of Facebook issuing an advisory opinion on the compatibility of digital nudging with the right to freedom of thought. This body has the ear of one of the most influential companies on the planet, and has the opportunity to clearly articulate whether the measures being considered and adopted by Facebook to manage misinformation are aligned with the human right to freedom of thought. Such work can also inform the related practices of other social media platforms. Another step is the UN Human Rights Committee issuing a General Comment on the right to freedom of thought, independent of its manifestation and links to other rights, and specifically on its applicability and relevance when people are subjected to the digital realm, while considering social media use and its impact. Such steps would be welcome and assist other law and policymakers and shapers within states in their own bottom-up approaches to governance.

For now, just because the law regarding freedom of thought does not currently prevent digital nudges from being further implemented on social media newsfeeds, does not automatically mean such a measure should be used without further debate, even if it is for the purpose of addressing misinformation and its spread. Conduct can be lawful but wrong. In order to avoid falling into this category, the risks and shortcomings of digital nudging, which extend beyond those raised here, need to be considered and acted upon. Nudging raises flags when reflecting on freethinking and choice, and the nudge may yet be proven to be incompatible with the rights

associated with such processes of the human mind. The nudge paradox highlights as much.

Social media platforms' content curation and moderation practices are also not enough to dissipate misinformation, even if they help manage it. Digital nudges have the potential to reduce the spread of misinformation on these platforms and may well be the least rights-intrusive means of doing so when compared to alternatives (with the possible exception of user-reporting). However, they are nonetheless a limited means of addressing misinformation because they can only manage the symptoms of this issue. Treating the causes involves having and acting on uncomfortable conversations about how companies and states are governed, and who that governance actually benefits. Reliance on digital nudging to solve societal issues associated with misinformation tolerates the maintenance of modes of operation that arguably require changing. A pertinent example is that this measure does not reduce user-engagement on social media, but in part relies on it. Utilizing digital nudges are grounds for cautious optimism in the effort to effectively address misinformation and its spread whilst respecting freedom of thought. Yet human rights and the laws enshrining them should form the cornerstone of decisions about whether and how digital nudging is continued, expanded, limited, or prohibited on social media newsfeeds.