DISENTANGLING DISINFORMATION: WHAT MAKES REGULATING DISINFORMATION SO DIFFICULT?

Jason Pielemeier*

I. INTRODUCTION

Since the 2016 U.S. elections, online disinformation has joined hate speech, terrorist incitement, and other forms of "harmful online content" as a key target for corporate and government policy makers.[1] Most major content platforms have developed policies and other approaches for disinformation, while legislative and regulatory proposals specifically designed to address online disinformation have been enacted in consolidated democracies, like France, unconsolidated democracies, like Malaysia, and autocratic states, like Singapore.[2] Meanwhile, several other jurisdictions have begun considering proposals to address disinformation, together with other content issues, through a single, comprehensive regulatory framework.[3] These laws and other similar proposals have sparked considerable debate, with critics focusing primarily on the effects—whether intended or not—that such measures could have on freedom of opinion and expression.[4]

[*] © 2020 Jason Pielemeier. Policy Director at the Global Network Initiative (GNI). This Essay is written in his personal capacity and does not represent the views of the GNI or any of its members.

[1] *See infra* Part 25

[2] *See infra* Sections IV.B.1, IV.B.2 (discussing the E.U. and France's laws regarding disinformation and the manipulation of information).

[3] The Government of the United Kingdom released a comprehensive proposal to address "online harms" by empowering a regulatory authority to develop issue-specific "codes" and enforcing a broad "duty of care" on a large swath of digital service providers. *See generally* SECRETARY OF STATE FOR DEPARTMENT OF DIGITAL, CULTURE, MEDIA & SPORT & SECRETARY OF STATE FOR THE HOME DEPARTMENT, ONLINE HARMS WHITE PAPER, 2019, CP 57, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf [https://perma.cc/BHR8-DAU5] (releasing the United Kingdom's proposal for addressing online harms). The Irish government has proposed a similar approach in its "Online Safety and Media Regulation Bill 2019." *See General Scheme Online Safety Media Regulation Bill 2019*, GOV'T OF IR., DEP'T OF COMM., CLIMATE ACTION, & ENV'T, https://www.dccae.gov.ie/en-ie/communications/legislation/Pages/General-Scheme-Online-Safety-Media-Regulation.aspx [https://perma.cc/E2PC-2TB5] (last visited May 26, 2020). Most recently, the new President of the European Commission has announced plans to develop a comprehensive, European Union-wide "Digital Services Act." *See* Kenneth Propp, *The Emerging EU Regulatory Landscape for Digital Platform Liability*, ATL. COUNCIL (Oct. 22, 2019), https://www.atlanticcouncil.org/blogs/new-atlanticist/the-emerging-eu-regulatory-landscape-for-digital-platform-liability/ [https://perma.cc/K42E-9KBZ].

[4] *See, e.g.*, U.N. Special Rapporteur on Freedom of Op. and Expression, Org. for Sec. & Co-operation in Eur. Representative on Freedom of the Media, Org. of Am. States Special

This Essay articulates some of the critical ways in which disinformation differs from other categories of harmful content and explores some of the early efforts by platforms and governments to address the issue. It begins by analyzing the semantics around disinformation, explaining how specific terminology can allude to distinct concerns. It then explores the similarities and differences between disinformation and related categories of harmful content, like hate speech and terrorist incitement, before examining some of the corporate and regulatory initiatives that have emerged. It concludes with some observations and cautionary notes for corporate and governmental policy makers as they consider how best to address disinformation.

II. DEFINING DISINFORMATION

There are a variety of terms used to describe the ways information is (mis)used to shape people's beliefs and behavior.[5] "Disinformation" has emerged as the most popular term used by government regulators to broadly describe the kinds of online-specific manipulation that they are most concerned about. Although there is a plethora of definitions of "disinformation," in this Essay, I will use the European Commission's definition from the "Communication on tackling online disinformation" and "Action Plan Against Disinformation."[6]

Rapporteur on Freedom of Expression, & African Comm'n on Human & Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Info., Joint Declaration on Freedom of Expression and "Fake News," Disinformation and Propaganda, U.N. Doc. FOM.GAL/3/17, at 1 (March 3, 2017), https://www.osce.org/fom/302796?download=true [https://perma.cc/ML3K-Z9K6] [hereinafter Joint Declaration] ("*Stressing* that the human right to impart information and ideas is not limited to 'correct' statements, that the right also protects information and ideas that may shock, offend and disturb.") (emphasis in original).

[5] For a helpful primer on these different terms, see generally Dean Jackson, *Issue Brief: Distinguishing Disinformation from Propaganda, Misinformation, and "Fake News*,*"* NAT'L ENDOWMENT FOR DEMOCRACY (Oct. 17, 2017), https://www.ned.org/issue-brief-distinguishing-disinformation-from-propaganda-misinformation-and-fake-news/ [https://perma.cc/8BXU-F52H] (defining the term disinformation and explaining why the current information environment amplifies disinformation). *See also* CAROLINE JACK, DATA & SOCIETY, LEXICON OF LIES: TERMS FOR PROBLEMATIC INFORMATION 2–8, 11–12 (Aug. 9, 2017), https://datasociety.net/pubs/oh/DataAndSociety_LexiconofLies.pdf [https://perma.cc/87YY-LWY3] (explaining the differences between the terms propaganda, disinformation, and misinformation); Claire Wardle & Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, COUNCIL OF EUR. 4, 20–42 (Sept. 27, 2017), https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77 [https://perma.cc/YJ48-RSLE] (examining "information disorder" and defining terms to "capture the complexity of the phenomenon").

[6] *See generally Commission Communication for Tackling Online Disinformation: A European Approach*, COM (2018) 236 final (Apr. 26, 2018), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0236&from=EN [https://perma.cc/7BJW-5ZTS] [hereinafter *Tackling Online Disinformation*] (discussing the EU's approach to

As I explain later in this Essay,[7] that definition is a variation on an earlier one developed by a High Level Expert Group ("HLEG") convened by the Commission to advise on policy initiatives concerning those topics, and thus has a degree of multi-stakeholder validation and purchase that many other definitions lack.[8] The Commission's definition is also quite broad and is therefore likely to include most, if not all, of the kinds of content that lawmakers in different contexts are concerned about (for instance, some governments are focused primarily on foreign propaganda, while others are more concerned about economically motivated disinformation, and the EU's definition covers both). For that same reason, it may also include some categories of information that are difficult to distinguish from speech that is traditionally considered protected. As I discuss below, that critique helps illustrate one of the important ways in which efforts to address disinformation through regulation may need to differ from attempts to address other forms of online content.

The Commission defines disinformation as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm."[9] As the table below illustrates, this definition covers deliberately spreading false news (often referred to as "fake news"), marketing products using false information, and distributing altered content (modified records, deep fakes, deceptively edited content or "shallow fakes," etc.).[10] In some contexts, it may also include "blended" information (with elements of true and false content) and true information that is propagated with the intent to deceive the public, which is sometimes described as "propaganda" or "mal-information."

However, the Commission was deliberately ambiguous about its relationship to categories of content that are already illegal (defamation, hate speech, etc.) and made it clear that the definition does not include objectively false information that is spread by those who believe it to be true (or are uncertain about its veracity), which is often referred to as "misinformation," nor "reporting errors, satire and parody,

tackling online disinformation); EUR. COMM'N, EU CODE OF PRACTICE ON DISINFORMATION (2018), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454 [https://perma.cc/W3T2-ZRKF] [hereinafter EU CODE OF PRACTICE] (outlining the EU's code for regulating disinformation).

[7] *See infra* Section IV.B.55.

[8] *See generally* Indep. High Level Group on Fake News and Online Disinformation, A Multi-Dimensional Approach to Disinformation (March 2018), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271 [https://perma.cc/D7ZP-3R3D] [hereinafter High Level Group Report] (reporting to the European Commission that, given the complexity of the disinformation issue, a multi-shareholder solution is required).

[9] *Tackling Online Disinformation*, *supra* note 6, at § 2.1.

[10] For more on the distinction between deep fakes, shallow fakes, and other forms of media manipulation, see *Deepfakes, Shallowfakes and Speech Synthesis: Tackling Audiovisual Manipulation*, Eur. Parliamentary Research Serv. (Dec. 4, 2019), https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/ [https://perma.cc/KC5W-XFA6].

[and] clearly identified partisan news and commentary."[11] It also arguably excludes opinion or true information that is propagated with the intent to distract or influence, which is sometimes described as "strategic communication."[12] Other categories of content that appear not to be included under this definition include: doxxing (disseminating identifying or private information about an individual or organization), marketing, and predictions.

Some companies, researchers, and advocates have focused less on what constitutes disinformation and more on how inauthentic information is spread.[13] This is in part because it can be easier to determine when certain methods and technical tools of dissemination, including bots, paid amplification, and/or coordinated campaigns, are used than it is to evaluate the veracity of the underlying content.[14] This approach tends to prioritize the scale of the campaign over the potential harm of the information being shared.

Dissecting the EU's Definition of Disinformation

Definitional components			**Examples**
Verifiably false information created, presented, and disseminated	. . . for economic gain	. . . may cause public harm.	For-profit, anti-vax campaigns; Click-bait targeted along racial, ideological, or other politically-motivated themes
		. . . does not cause public harm.	*Spurious click-bait*
	. . . to intentionally deceive the public	. . . may cause public harm.	False news or deep fakes generated to spur hatred, division, political goals, etc.
		. . . does not cause public harm.	*Satire, parody, or comedy.*
Misleading information created, presented, and disseminated	. . . for economic gain	. . . may cause public harm.	General or targeted marketing, search-engine optimization, or promotion of products that can cause public harm (e.g., counterfeit pharmaceuticals).
		. . . does not cause public harm.	*General or targeted marketing, search-engine optimization, or*

[11] EU CODE OF PRACTICE, *supra* note 6, at Preamble.

[12] *See generally* Kirk Hallahan et al., *Defining Strategic Communication*, 1:1 INT'L J. STRATEGIC COMM. 3 (Mar. 2007) (examining the nature of strategic communication through six relevant disciplines).

[13] *See infra* Section IV.A.39.

[14] *See* Camille François, Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses 2–6 (Sept. 20, 2019) (working paper) (on file with the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression), https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf [https://perma.cc/ECH4-BVAD].

			promotion of natural supplements/ remedies, consumer goods, etc.; "reporting errors"
	. . . to intentionally deceive the public	. . . may cause public harm.	Shallow fakes; opinion/true information targeted at particular groups to spur hatred, division, political goals, etc.
		. . . does not cause public harm.	*Shallow fakes or opinion with minimal impact*
Other excluded content		*Illegal content that is not false or misleading*	*Hate speech, defamation, incitement to violence, revenge-porn, etc.*
		"Clearly identified partisan news and commentary"	*Campaign ads and other political content*
		True information, even if shared with malicious intent	*Doxxing; unaltered intimate images; unaltered hacked materials (if not misleading)*

III. WHAT MAKES DISINFORMATION DIFFERENT?

Disinformation shares certain key characteristics with other categories of harmful content, such as terrorist incitement and hate speech. All of these categories are notoriously difficult to define in ways that are consistent with freedom of expression, and all of them are difficult to apply in instances where context is limited or difficult to objectively discern. In addition, they each tend to produce diffuse social impacts, rather than narrow personal harms, thereby minimizing the extent to which states or platforms can rely on "victims" to assist in identifying and enforcing relevant rules. Finally, in the online context, the quantity of content that arguably falls into these categories is staggering, and its sources may be anonymous/pseudonymous and dispersed (including across platforms and borders), making it difficult, if not impossible, for any one nation-state or platform to effectively address it in isolation.

These similarities, taken together, may explain the impulse on the part of some states to attempt to regulate the identification and enforcement of these categories under a single, comprehensive, regulatory framework. Notwithstanding these commonalities, disinformation has three distinguishing characteristics that make it uniquely difficult for nation-states and platforms to effectively address, especially under a single, catch-all approach.

A. Broader Spill-Over Effects (the "Definition Challenge")

First, while all categories of harmful content—including disinformation, terrorist incitement, and hate speech—are challenging to define, disinformation is especially difficult to prohibit without causing potentially broad impacts on lawful and protected speech. Although it can be difficult to draw the line between speech that is hateful but lawful and hate speech, or content that celebrates terrorist causes and that which incites terrorist acts, many countries have nevertheless drawn these lines in ways that survive scrutiny under domestic and international human rights law—that is to say, in clear, narrow terms that are justified, necessary, and proportionate.[15] By comparison, it can be extremely difficult to objectively determine the truth in a given context, much less establish whether an individual knew or should have known that certain information was untrue or misleading. This is a challenge that also arises in the application of reputation protection (i.e., defamation) laws, where truth is often considered an affirmative defense against liability.

Another way to frame this is to say that the quantity of information that could be reasonably mistaken for hate speech or terrorist incitement is substantially smaller than that which could be confused for disinformation. For similar reasons, the potential for authorities to willingly misapply prohibitions on disinformation to censor and punish protected speech is arguably greater than it is vis-à-vis hate speech or terrorist incitement (although those categories have both been famously abused in the past[16]).

Finally, even where disinformation laws are carefully drafted and applied, their potential chilling effects (i.e., likelihood of causing preemptive self-censorship) can nevertheless be significant since they could cause individuals to refrain from sharing content (as well as opinions) they perceive as objective or newsworthy but cannot independently or reliably verify. Since the public interest value of information that might be mistaken for disinformation (e.g., political opinion or independent journalism) can be much higher than legal but hateful content or terrorist propaganda, the ramification of such chilling effects in the context of disinformation are arguably much greater.

B. The Challenges of Proving Intent (the "Intent Challenge")

Second, unlike many prohibitions on hate speech and terrorist incitement, proving disinformation generally requires establishing intent on the part of the

[15] *See, e.g.*, Joint Declaration, *supra* note 4, at 2–3. Indeed, while it is relatively easy to identify which of the enumerated "legitimate purposes" justify prohibitions on hate speech and terrorist incitement, it is much harder to articulate which purpose or combination of purposes are served by prohibiting disinformation.

[16] *See, e.g.*, LEWIS GORDON ET AL., OAKLAND INST., ETHIOPIA'S ANTI-TERRORISM LAW: A TOOL TO STIFLE DISSENT 9–11 (2015), https://www.oaklandinstitute.org/sites/oaklandinstitute.org/files/OI_Ethiopia_Legal_Brief_final_web.pdf [https://perma.cc/E9FB-CSZE]; *The Global Gag on Free Speech Is Tightening: In Both Democracies and Dictatorships, It Is Getting Harder to Speak Up*, ECONOMIST (Aug. 17, 2019), https://www.economist.com/international/2019/08/17/the-global-gag-on-free-speech-is-tightening [https://perma.cc/R8FV-Q72D] (describing how hate speech and other laws are misused for censorship).

speaker.[17] Determining a speaker's intent is notoriously difficult[18] and can be doubly difficult in online contexts where nuance, jargon, and slang—not to mention the use of different languages—proliferate.

This challenge is compounded by the fact that disinformation, by definition, often must also have the potential to cause "public harm."[19] This implication of seriousness and scale suggests that, in many instances, a large number of individuals have spread the disinformation, despite the fact that they may not share the same intent to deceive. In other words, even if the intent of the author can be established, it may still be near impossible to prove the intent of others who have subsequently shared her disinformation. For this reason, some efforts to address disinformation have emphasized "traceability"—the ability to identify where information originated and has since spread—in a manner that laws addressing hate speech and terrorist incitement have not.[20] This Essay explores the privacy implications of such requirements further below.

Notably, the Commission's definition bifurcates "economic gain" and "intent to deceive," which could be read to prohibit disinformation spread with a remunerative goal, even where there is no demonstrated intent to deceive, or even awareness that the underlying information is misleading or inaccurate. While this broad approach could ease enforcement efforts to some degree, it could also increase the potential for over-enforcement and the chilling of certain economic activity.[21]

C. Diffuse and Lasting Impacts (the "Harm Challenge")

Third, whereas the very purpose of hate speech and terrorist incitement is to target specific individuals or groups by instilling fear and/or motivating actual harm, the purpose of any given piece of disinformation may be less clear and its impacts more diffuse. This is especially true of individual instances of disinformation that cannot be conclusively tied to broader campaigns. This makes it difficult to

[17] *See* High Level Group Report, *supra* note 8 and accompanying text (discussing European Commission's definition of disinformation).

[18] *See generally* Leslie Kendrick, *Speech, Intent, and the Chilling Effect*, 54 WM. & MARY L. REV. 1633 (2013) (discussing how a speaker's intent is difficult to determine and how the chilling effect may not be a justification for speaker's intent requirements).

[19] High Level Group Report, *supra* note 8, at 5.

[20] *See infra* note 79 and accompanying text (explaining Singapore's implicit requirement that internet intermediaries employ methods to trace user activity). The Indian government has also proposed a set of draft rules that would require traceability. *See* Ministry of Electronics & Info. Tech., The Information Technology [Intermediaries Guidelines (Amendment) Rules] 3(5) (Dec. 24, 2018) (India), https://meity.gov.in/writereaddata/files/Draft_Intermediary_Amendment_24122018.pdf [https://perma.cc/N6TL-S3PX] (requiring that intermediaries "shall enable tracing out of such originator of information on its platform as may be required by government agencies who are legally authorised").

[21] For instance, journalistic publication is often done for a variety of reasons, including economic gain.

objectively establish and measure harm, thereby making it difficult to enforce prohibitions where evidence of large-scale, coordinated propagation is lacking.[22] In addition, where individuals or entities are targeted for enforcement, they will often be able to justifiably complain about selective enforcement.

To the extent that large-scale disinformation efforts may rely on a combination of inauthentic and/or boosted dissemination, as well as organic and uncoordinated amplification, it will also be difficult to identify precisely how much blame to attribute to the former versus the latter. Likewise, when disinformation campaigns exploit existing social divisions or lack of public awareness,[23] it may be challenging to isolate and measure their "public harm." As a result, the impacts and echoes of disinformation may linger long after particular content is identified and removed.

Because it can take longer to identify disinformation, the perceived need to "remedy" the harm it creates and "correct the record" can drive an understandable impulse to track disinformation, not only horizontally (as it spreads to new users in real time) and prospectively (through filtering), but also retroactively. Horizontal and prospective tracking can be accomplished using metadata, data "hashes," and even random "spot checks."[24] However, since it is impossible to know *ex ante* who will propagate disinformation, retrospective tracking requires the development of capabilities that would presumably impact all users and could result in *de facto* prohibitions on the use of encryption.[25] Even if these capabilities are used narrowly in specific, pre-defined, and carefully overseen circumstances to address disinformation, they would create a universe of possibility that would be very hard for law enforcement to ignore and would almost certainly generate chilling effects.[26]

In sum, all the elements that are traditionally necessary to prove a violation of criminal law—*actus reas*, *mens rea*, and damages—are more difficult to establish

[22] *See infra* note 41 and accompanying text.

[23] *See, e.g.*, Adam Entous et al., *Russian Operatives Used Facebook Ads to Exploit America's Racial and Religious Divisions*, WASH. POST (Sept. 25, 2017, 03:15 PM MDT), https://www.washingtonpost.com/business/technology/russian-operatives-used-facebook-ads-to-exploit-divisions-over-black-political-activism-and-muslims/2017/09/25/4a011242-a21b-11e7-ade1-76d061d56efa_story.html [https://perma.cc/U8MA-N2WT] (pointing out that disinformation campaigns attributed to the Russian government have sought to exacerbate racial and religious divisions in the U.S.).

[24] For recent analyses of efforts along these lines, see Amelia Acker, *Tracking Disinformation by Reading Metadata,* MEDIUM (July 17, 2018), https://medium.com/@MediaManipulation/tracking-disinformation-by-reading-metadata-320ece1ae79b [https://perma.cc/V5SX-LZHX] and GLOBAL DISINFORMATION INDEX, https://disinformationindex.org/the-index/ [https://perma.cc/J8KZ-S6DQ] (last visited Apr. 8, 2020).

[25] *See* Jennifer Daskal, *This 'Fake News' Law Threatens Free Speech. But it Doesn't Stop There.: Singapore's New Legislation Could Force Companies to Tell the Government What Websites Users Have Viewed.*, N.Y. TIMES (May 30, 2019), https://www.nytimes.com/2019/05/30/opinion/hate-speech-law-singapore.html [https://perma.cc/7TAP-BAWX] (discussing the difficulties of enforcing a disinformation law).

[26] *See id.*

vis-à-vis disinformation than they are in the context of hate speech, terrorist incitement, or most other categories of harmful content. As a result, it is not surprising that efforts to-date to address disinformation have differed in important ways.

IV. HOW ARE COMPANIES AND GOVERNMENTS ADDRESSING DISINFORMATION?

Efforts by governments and platforms to address hate speech and terrorist incitement have focused primarily on detecting and removing offending content as quickly as possible. However, because of how difficult it can be to define and apply prohibitions on disinformation, as well as legitimate concerns about platforms becoming arbiters of truth,[27] governments and platforms have tended to propose solutions short of censorship. This Part provides an overview of company and government efforts to address disinformation. As this overview makes clear, there is a strong interplay between these actors, with governments pushing companies to take voluntary actions while threatening to provide greater oversight and accountability through regulation if necessary.

A. Company Efforts

Large content platforms have taken a variety of steps to address disinformation and these efforts have been influenced by a wide range of both regulatory and non-regulatory factors. In general, company efforts can be grouped into five categories: (1) limiting the reach of false news/information; (2) demonetization; (3) addressing inauthentic behavior; (4) contextualization; and (5) transparency. These efforts often allow the offending content to remain visible, albeit sometimes to a smaller audience and/or with signals or context that can help users understand the contested nature of the information being presented.

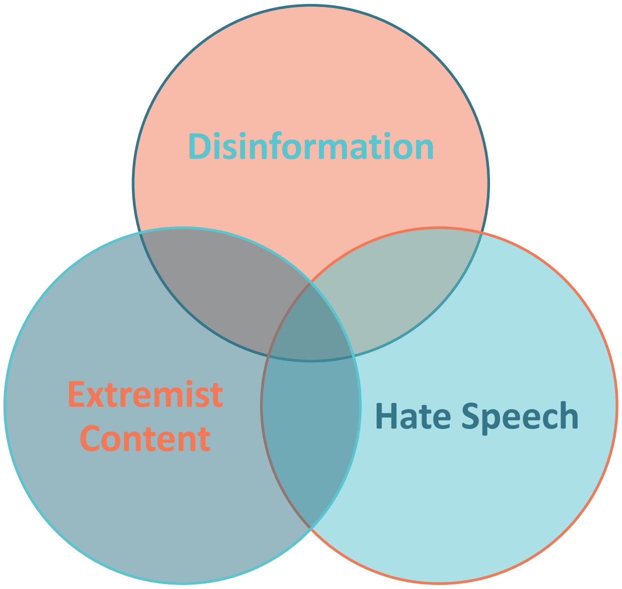
It should be noted that, while almost all platforms have refused to prohibit fake news or disinformation as such,[28] as the Venn diagram below illustrates, disinformation often contains elements of hate speech or extremist content that can

[27] *See* Tony Romm, *Zuckerberg: Standing for Voice and Free Expression*, WASH. POST (Oct. 17, 2019, 02:22 PM MDT), https://www.washingtonpost.com/technology/2019/10/17/zuckerberg-standing-voice-free-expression/ [https://perma.cc/T8FF-2C2E] (transcribing Mark Zuckerberg's speech, "Standing for Voice and Free Expression: Speech Delivered at Georgetown University").

[28] *See* John Oates, *So How Well Did You Block Fake News, Google? Facebook? Web Goliaths Turn in Self-Assessment Homework to Europe: Unsurprisingly, Commish Thought That, in Fact, They Could Do Better*, REGISTER (Oct. 30, 2019, 11:19), https://www.theregister.co.uk/2019/10/30/eu_first_reports_on_disinformation_from_google_twitter_and_facebook/ [https://perma.cc/K3QS-HKJAPERMA] (discussing Facebook and Google's internal policies regarding the removal of disinformation).

be and often are otherwise targeted for removal.[29] In addition, it is important to recognize that smaller platforms often lack the resources required to implement many of the systemic and nuanced policies set out below. As a result, some may have to rely more on users/community moderators (as in the case of Reddit[30]) or choose a more black-and-white approach (as Pinterest has done[31]).

Overlaps Between Categories of "Harmful Content"



[29] Disinformation can also incite violence in ways that would violate platform rules. *See*, *e.g.*, *Facebook Community Standards: 1. Violence and Incitement*, FACEBOOK, https://www.facebook.com/communitystandards/credible_violence [https://perma.cc/4VFT-Z6YJ] (last visited Mar. 25, 2020) ("While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence.").

[30] *See, e.g.*, *Managing Misinformation on Reddit*, NAT'L PUB. RADIO (Dec. 8, 2019, 8:01 AM), https://www.npr.org/2019/12/08/786039738/managing-misinformation-on-reddit [https://perma.cc/M52N-EGTA] (explaining how Reddit relies on volunteer moderators).

[31] *Community Guidelines*, PINTEREST, https://policy.pinterest.com/en/community-guidelines [https://perma.cc/6NEK-QB2H] (last visited Apr. 10, 2020) (setting out a policy that prohibits "[c]ontent that originates from disinformation campaigns").

1. *Limiting Reach*

Major platforms have focused on improving the speed and accuracy with which they can identify disinformation peddled on their platforms by, for instance, improving machine learning detection tools and building better ways to work with users and third parties to identify and prevent it from spreading on their platforms.[32] Once disinformation is identified, platforms can limit its spread by, among other things, altering its ranking in algorithmic feeds in order to reduce its prevalence.[33] In addition to demoting false information, platforms have also taken steps to promote verified, authentic information and reporting, especially in the context of breaking news.[34] This practice can have a similar and complementary effect as demotion by effectively making disinformation harder to access.

2. *Demonetization*

Platforms have also attempted to limit the extent to which purveyors of disinformation can use tools designed for advertising and marketing (e.g., purchasing and targeting ads, "boosting" content) to augment their reach.[35] For instance, Facebook has said that it is "[m]aking it as difficult as possible for people

[32] *See* Donald Hicks & David Gasca, *A Healthier Twitter: Progress and More to Do*, TWITTER: BLOG (Apr. 16, 2019), https://blog.twitter.com/en_us/topics/company/2019/health-update.html [https://perma.cc/6SBG-C5KB] (heralding the use of technology to track spam, platform manipulation, and other rule violations); Adam Mosseri, *Working to Stop Misinformation and False News*, FACEBOOK FOR MEDIA (Apr. 7, 2017), https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news [https://perma.cc/R43T-NYAD] (stating Facebook is applying machine learning to assist response teams); Charlie Warzel, *Why Can Everyone Spot Fake News But YouTube, Facebook, and Google?*, BUZZFEED NEWS (Feb. 22, 2018, 7:40 PM), https://www.buzzfeednews.com/article/charliewarzel/why-can-everyone-spot-fake-news-but-the-tech-companies [https://perma.cc/QCK3-FU8R] (citing YouTube stating that "it uses machine learning to flag possibly violative content for human review").

[33] *See* Mosseri, *supra* note 32; *Continuing Our Work to Improve Recommendations on YouTube*, YOUTUBE: OFFICIAL BLOG (Jan. 25, 2019), https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html [https://perma.cc/QXB9-W3CG] (promising to "tak[e] a closer look at how we can reduce the spread of content that comes close to—but doesn't quite cross the line of—violating our Community Guidelines").

[34] *See, e.g.*, *Breaking New and Top News on YouTube*, YOUTUBE: HELP, https://support.google.com/youtube/answer/9057101?hl=en [https://perma.cc/LT7R-NNCU] (last visited Apr. 10, 2020) (showing YouTube's "Breaking News" feature as an example of a platform which has taken steps to promote authentic information).

[35] For a summary of the ways that disinformation campaigns (ab)use the features inherent to digital, advertising-driven business models, see DIPAYAN GHOSH & BEN SCOTT, NEW AMERICA, #DIGITALDECEIT: THE TECHNOLOGIES BEHIND PRECISION PROPAGANDA ON THE INTERNET, (Jan. 2018), https://www.newamerica.org/documents/2077/digital-deceit-final-v3.pdf [https://perma.cc/C2MH-PJBC].

posting false news to buy ads on our platform through strict enforcement of our policies,"[36] while Google has enhanced its efforts to address misinformation placed in Google Ads, as well as through its "AdSense" service that helps publishers fund their own content by placing ads on their websites.[37] For its part, Twitter has developed an "unacceptable business practices" ads policy, which prohibits advertising for accounts making misleading, false or unsubstantiated claims,[38] as well as a "quality policy," through which ads are reviewed to ensure they adhere to editorial guidelines.[39]

3. *Addressing Inauthentic Behavior*

In recent months, platforms have also increased attention to the ways that fake accounts, spam, and "inauthentic" behavior are used as part of disinformation campaigns. This includes technical measures to detect the use of bots to create or coordinate accounts, as well as efforts to artificially "optimize" engagement or otherwise manipulate algorithmic feeds, including news feeds and search engine results.[40] For instance, Twitter's new "Platform Manipulation and Spam Policy" prohibits a range of behaviors, including: commercially-motivated spam; inauthentic engagements "that attempt to make accounts or content appear more popular or active than they are;" and coordinated activity "that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting."[41] Given Facebook's somewhat unique "real name" policy,[42] which essentially prohibits pseudonymous accounts, it has been able to more specifically define and enforce "inauthentic behavior."[43]

[36] Mosseri, *supra* note 32.

[37] *See* GOOGLE, EC EU CODE OF PRACTICE ON DISINFORMATION: GOOGLE ANNUAL REPORT 5–7 (Oct. 29, 2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62680 [https://perma.cc/HJR4-VCA4] [hereinafter, GOOGLE DISINFORMATION REPORT].

[38] *See Unacceptable Business Practices*, TWITTER: BUSINESSES, https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/unacceptable-business-practices.html [https://perma.cc/YES2-TUGV] (last visited Apr. 10, 2020) (stating Twitter's ad policies for business practices).

[39] *Quality Policy*, TWITTER: BUSINESSES, https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/Quality_Policy.html [https://perma.cc/5XZE-MPR4] (last visited Apr. 10, 2020) (stating Twitter's quality policy regarding ads).

[40] This is what Google refers to as "Engagement Abuse." *See* GOOGLE DISINFORMATION REPORT, *supra* note 37, at 16.

[41] *Platform Manipulation and Spam Policy*, TWITTER: HELP CTR. (Sept. 2019), https://help.twitter.com/en/rules-and-policies/platform-manipulation [https://perma.cc/3R23-6UQS].

[42] *See What Names Are Allowed on Facebook?*, FACEBOOK, https://www.facebook.com/help/112146705538576 [https://perma.cc/4M3T-US7E] (last visited Apr. 10, 2020) (stating the names users cannot use and other things to keep in mind).

[43] Facebook defines "inauthentic behavior" as the:

4. Contextualization

Recognizing that disinformation will continue to exist on these platforms, notwithstanding efforts to limit its spread, platforms have also been providing additional resources to users in an attempt to limit the pernicious impacts of disinformation they may encounter. These efforts include providing additional context around content that may constitute disinformation, fact checking, and media literacy campaigns. For example, Facebook says it has been "exploring ways to give people more context about stories so they can make more informed decisions about what to read, trust, and share and ways to give people access to more perspectives about the topics that they're reading."[44] This includes partnering with local NGOs to provide digital skills training and education,[45] working with newsrooms and journalists,[46] and collaborative efforts with researchers working on disinformation, including providing access to Facebook data and funding research.[47] Meanwhile, Google has developed tools like the "Share the Facts" widget to facilitate fact checking and partnered with media to develop signals of trustworthiness, as well as content and source credibility.[48] Smaller platforms have also engaged in user literacy efforts. For instance, the social media platform Tumblr recently rolled out six

> [U]se of Facebook or Instagram assets (accounts, pages, groups, or events), to mislead people or Facebook: about the identity, purpose or origin of the entity that they represent; about the popularity of . . . content or assets; about the purpose of an audience or community; about the source or origin of content; [or] to evade enforcement under . . . Community Standards.

Facebook Community Standards: 20. Inauthentic Behavior, FACEBOOK, https://www.facebook.com/communitystandards/inauthentic_behavior [https://perma.cc/5Z85-R54K] (last visited Apr. 10, 2020). "Coordinated Inauthentic Behavior" is defined as "the use of multiple . . . assets, working in concert to engage in Inauthentic Behavior," including on behalf of a government actor. *Id.*

[44] Mosseri, *supra* note 32.

[45] *See Digital Literacy Library*, FACEBOOK, https://www.facebook.com/safety/educators [https://perma.cc/G3NQ-E9S6] (last visited Apr. 10, 2020).

[46] *See Welcome to the Facebook Journalism Project*, FACEBOOK: JOURNALISM PROJECT, https://www.facebook.com/journalismproject [https://perma.cc/RB3N-W4G6] (last visited Apr. 10, 2020); *Introducing the News Integrity Initiative*, FACEBOOK FOR MEDIA (Apr. 2, 2017), https://www.facebook.com/facebookmedia/blog/introducing-the-news-integrity-initiative [https://perma.cc/L3DN-WAHG]; NEWS LITERACY PROJECT, http://www.thenewsliteracyproject.org/ [https://perma.cc/BK4W-7Y4B] (last visited Apr. 10, 2020).

[47] FACEBOOK, FACEBOOK REPORT ON IMPLEMENTATION OF THE CODE OF PRACTICE FOR DISINFORMATION: ANNUAL REPORT SEPTEMBER 2019 5.1–.2 (Oct. 29, 2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62681 [https://perma.cc/8EUC-2Z5A] [hereinafter FACEBOOK: CODE OF PRACTICE].

[48] GOOGLE DISINFORMATION REPORT, *supra* note 37, at 17–19.

educational videos targeted at their users, including one on fake news and another on "Authenticity Online," using GIFs, short texts, and memes.[49]

5. Transparency

In addition to providing context and encouraging critical thinking, the platforms have also been expanding transparency around their efforts to enforce the above-mentioned platform rules. For instance, Facebook's most recent *Community Standards Enforcement Reports* ("CSER"), include metrics for its enforcement of policies against "fake accounts" and "spam."[50] Facebook has also created a Data Transparency Advisory Group, "made up of international experts in measurement, statistics, criminology and governance" tasked with providing an independent, public assessment of whether the metrics used in the CSER provide accurate and meaningful measures of Facebook's content moderation challenges and efforts to address them.[51] Google has also begun reporting on the number of removals for "spam, misleading content, and scams" on YouTube. According to its most recent transparency report, such content accounted for 58.9% of all removals on that platform.[52]

Facebook, Google, and Twitter have also clarified their political advertising policies, creating public repositories of such ads, and increasing reporting on their efforts to enforce their policies (although this has been an area of much controversy).[53] Meanwhile, Mozilla—the company that develops the open-source web browser Firefox—has developed a model "effective ad archive" API (Application Programing Interface) and challenged the larger platforms to ensure that their own ad APIs meet that standard.[54]

[49] Julia Alexander, *Tumblr Is Rolling Out an Internet Literacy Initiative to Help Combat Misinformation and Cyberbullying*, THE VERGE (Jan. 6, 2020, 10:00 AM), https://www.theverge.com/2020/1/6/21048134/tumblr-misinformation-2020-election-cyberbullying-digital-literacy [https://perma.cc/LAA9-5SBW].

[50] *See* FACEBOOK, COMMUNITY STANDARDS ENFORCEMENT REPORT (Nov. 2019), https://transparency.facebook.com/community-standards-enforcement [https://perma.cc/BS3U-RS5L] (last visited March 26, 2020) (providing reporting on fake accounts and spam).

[51] Radha Iyengar Plumb, *An Independent Report on How We Measure Content Moderation*, FACEBOOK (May 23, 2019), https://about.fb.com/news/2019/05/dtag-report/ [https://perma.cc/C8VW-HRF8].

[52] *YouTube Community Guidelines Enforcement*, GOOGLE: TRANSPARENCY REPORT, https://transparencyreport.google.com/youtube-policy/removals [https://perma.cc/4W5V-EBBN] (last visited Apr. 10, 2020).

[53] *See, e.g.*, Elizabeth Culliford, *Factbox: How Social Media Sites Handle Political Ads*, REUTERS (Nov. 15, 2019, 11:37 AM), https://www.reuters.com/article/us-usa-election-advertising-factbox/factbox-how-social-media-sites-handle-political-ads-idUSKBN1XP22G [https://perma.cc/EV75-GPWZ] (discussing the growing pressure on online entities to stop carrying false or misleading political ads).

[54] *Facebook and Google: This Is What an Effective Ad Archive API Looks Like*, MOZILLA: BLOG (Mar. 27, 2019), https://blog.mozilla.org/blog/2019/03/27/facebook-and-

B. Governmental Efforts

A number of governments have developed approaches to disinformation online and many more are in the process of doing so. According to one source, at least 52 countries representing every region of the world have implemented or are actively considering some form of legal, regulatory, or policy approach to disinformation, misinformation, or fake news.[55] While a review of all these efforts is beyond the scope of this Essay, this Section provides analysis of four particular efforts that illustrate the diversity of approaches under consideration, as well as the potential challenges they each may raise.

1. EU Code of Practice on Disinformation

The European Commission has led the most coherent, coordinated, and sustained effort to address disinformation to date, working together with leading content platforms and advertising industry representatives. In late 2017, the Commission convened a *High Level Expert Group on Fake News and Online Disinformation*, which delivered a report in March of 2018 titled, "A Multi-Dimensional Approach to Disinformation."[56] That report informed the Commission's April 2018 Communication, which included a pledge to develop "an ambitious Code of Practice," building on the principles proposed by the High Level Expert Group and committing online platforms and the advertising industry to a range of objectives.[57]

The Code of Practice was published in September of 2018,[58] and a month later it was "signed" by Google, Mozilla, and Twitter, as well as by advertisers and advertising industry groups, each of which presented "roadmaps" detailing their respective plans "to extend their tools against disinformation to all EU Member States" (Facebook signed in November 2018 and Microsoft signed in May 2019).[59] It contains five core commitments related to: (1) scrutiny of ad placements; (2) political advertising and issue-based advertising; (3) integrity of services; (4) empowering consumers; and (5) empowering the research community,[60] as well as an "annex" listing "best practices" and examples of corresponding, existing

google-this-is-what-an-effective-ad-archive-api-looks-like/ [https://perma.cc/8CXU-ACQN].

[55] Daniel Funke & Daniela Flamini, *A Guide to Anti-Misinformation Actions Around the World*, POYNTER, https://www.poynter.org/ifcn/anti-misinformation-actions/ [https://perma.cc/S5BJ-SZGV] (last updated Apr. 9, 2018).

[56] *See* High Level Group Report, *s*upra note 8.

[57] *Tackling Online Disinformation*, *supra* note 6, at § 3.1.1.

[58] *See* EU CODE OF PRACTICE, *supra* note 6.

[59] *Roadmaps to Implement the Code of Practice on Disinformation*, EUR. COMM'N, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation [https://perma.cc/RBK8-36CE] (last updated Feb. 13, 2020).

[60] EU CODE OF PRACTICE, *supra* note 6, at II.A–E.

policies/actions by participating companies.[61] Notwithstanding all of this, members of the "Sounding Board" of the "Multistakeholder Forum on Disinformation Online" issued a statement criticizing the Code for "contain[ing] no common approach, no clear and meaningful commitments, no measurable objectives or KPIs, hence no possibility to monitor process, and no compliance or enforcement tool. . . ."[62]

As part of the Code, participating entities commit to producing annual reports detailing these efforts[63] and the Commission pledged to produce a report summarizing actions taken during the initial year of Code implementation ("assessment report").[64] In the interim, the Commission also asked Facebook, Google, and Twitter to provide monthly reports between January and May of 2019 detailing the steps they were taking to address disinformation in the context of the 2019 European Parliamentary elections.[65] After receiving the annual reports, the Commission issued a statement expressing mixed reviews and noting that, to date, actions to empower consumers and researchers lagged behind those related to advertising scrutiny, de-monetization, transparency, and integrity of services.[66] The Commission's full assessment report is expected to be produced in early 2020. In the Commission's words, "[s]hould the results under the Code prove unsatisfactory, the Commission may propose further measures, including of a regulatory nature."[67]

[61] *Id.* at Annex 2, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54455 [https://perma.cc/RXR7-5R39] (last visited Mar. 22, 2020).

[62] THE SOUNDING BRD., THE SOUNDING BOARD'S UNANIMOUS FINAL OPINION ON THE SO-CALLED CODE OF PRACTICE (Sept. 24, 2018), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54456 [https://perma.cc/UL26-N7AE] [hereinafter SOUNDING BOARD OPINION].

[63] *Annual Self-Assessment Reports of Signatories to the Code of Practice on Disinformation 2019*, EUR. COMM'N (last updated Oct. 31, 2019), https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019 [https://perma.cc/NF9H-TRKQ] (aggregating 2019 annual reports) [hereinafter *Annual Self-Assessment Report 2019*].

[64] *European Commission Contribution to the European Council: Action Plan Against Disinformation*, at 19, COM (2018) 36 final (Dec. 5, 2018), https://ec.europa.eu/commission/sites/beta-political/files/eu-communication-disinformation-euco-05122018_en.pdf [https://perma.cc/TEX3-6954].

[65] *See Last Intermediate Results of the EU Code of Practice Against Disinformation*, EUR. COMM'N (June 14, 2019), https://ec.europa.eu/digital-single-market/en/news/last-intermediate-results-eu-code-practice-against-disinformation [https://perma.cc/PQ7F-CFYW].

[66] *See Annual Self-Assessment Report 2019*, *supra* note 63.

[67] *Code of Practice on Disinformation One Year On: Online Platforms Submit Self-Assessment Reports*, EUR. COMM'N (Oct. 29, 2019), https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_19_6166 [https://perma.cc/5XQJ-HY4C].

2. *France's Law on Manipulation of Information*

Soon after his 2017 election, which featured allegations of foreign, state-sponsored hacking and disinformation, President Macron announced his intent to introduce a new law to address fake news in the context of elections.[68] The law that passed that year was eventually amended to focus more precisely on "manipulation of information,"[69] which it defined as "inaccurate or misleading allegations or imputations that falsely report facts, with the aim of changing the sincerity of an upcoming election" and provide a mechanism whereby individuals, public authorities, and political parties can seek expedited judicial review of certain content.[70] This mechanism only applies during the three months prior to certain voting events and to content that has been spread online "deliberately, artificially or automatically, and massively."[71] Where a judge finds that the content meets the definition and these criteria, they may take any and all "proportionate and necessary measures" to halt its dissemination.[72]

The law also sets out a "duty of cooperation" for online platforms, which includes the provision of tools for users to flag disinformation, as well as other transparency and media literacy commitments.[73] Finally, it gives new authorities to the Higher Audiovisual Council ("CSA") to oversee this "duty of cooperation," as

[68] Angelique Chrisafis, *Emmanuel Macron Promises Ban on Fake News During Elections*, GUARDIAN (Jan. 3, 2018, 3:45 PM), https://www.theguardian.com/world/2018/jan/03/emmanuel-macron-ban-fake-news-french-president [https://perma.cc/L4R2-MLL8].

[69] This change was reportedly made after the publication of an opinion by the French Constitutional Council in order to ensure that satire was not penalized under the law. *See La proposition de loi "Fake news" en partie réécrite en commission*, LE FIGARO (May 30, 2018), https://www.lefigaro.fr/flash-actu/2018/05/30/97001-20180530FILWWW00189-la-proposition-de-loi-fake-news-en-partie-reecrite-en-commission.php [https://perma.cc/QFL2-XWQJ].

[70] Proposition de Loi n° 190 du 20 novembre 2018 de relative à la lutte contre la manipulation de l'information [Proposal of Law No. 190 of November 20, 2018 relating to the fight against the manipulation of information], ASSEMBLÉE NATIONALE [NATIONAL ASSEMBLY], Nov. 20, 2018, Tit. I, Art. 1, http://www.assemblee-nationale.fr/15/ta/tap0190.pdf [https://perma.cc/NWC9-8Q3E] (amending the Electoral Code) [hereinafter Proposition de Loi n° 190].

[71] *Id.* at Tit. I, Art. 1, Art. L. 163-2.

[72] *Id.*

[73] *See id*. at Tit. III, Art. 11. *See also* Alexander Damiano Ricci, *French Opposition Parties Are Taking Macron's Anti-Misinformation Law to Court*, POYNTER (Dec. 4, 2018), https://www.poynter.org/fact-checking/2018/french-opposition-parties-are-taking-macrons-anti-misinformation-law-to-court/ [https://perma.cc/R5LV-CBJH] (discussing the duty of cooperation and how online platforms have to establish a "tool for users to flag disinformation" as well as other measures).

well as to revoke the radio and television broadcast rights of entities that disseminate disinformation and are "under the influence of" or "controlled" by a foreign state.[74]

3. Singapore's Protection from Online Falsehoods and Manipulation Act ("POFMA")

In May of 2019, Singapore enacted POFMA, the purpose of which is to: prevent communication of false statements; suppress online locations that repeatedly communicate false statements; enable measures to detect, control, and safeguard against coordinated inauthentic behavior; and enhance disclosure of information concerning paid political content.[75]

The law states circularly that "a statement is false if it is false or misleading" and provides significant criminal sanctions for anyone communicating such a statement in Singapore knowing or having reason to believe that it is false and "is likely to be" prejudicial in a variety of possible ways.[76] It also criminalizes knowingly providing services to be used in communicating offending content.[77] Uniquely, the law allows "any Minister" to order either a "correction direction" or a "stop communication direction" to the author, as well as to an internet intermediary, for any false statement of fact if they are "of the opinion that it is in the public interest."[78]

Notably, a targeted correction order can require an intermediary to communicate that correction "to all end-users in Singapore that it knows had accessed" the statement, implying that they must enable ways to track user activity retroactively.[79] Failure to comply with any such direction can result in criminal liability, as well as a blocking order.[80] All aspects of the law expressly apply extraterritorially.[81]

[74] Proposition de Loi n° 190, *supra* note 70, at Tit. II, Art. 10 (unofficial Google translation).

[75] *See* Protection from Online Falsehoods and Manipulation Act 2019 (No. 18 of 2019), June 3, 2019, pt. 1, § 5 (Sing.), https://sso.agc.gov.sg/Acts-Supp/182019/Published/20190625?DocDate=20190625 [https://perma.cc/KA2R-XPMT] [hereinafter, POFMA].

[76] *Id*. at pt. 1, § 2(2), pt. 2, § 7(1)(b).

[77] *See id*. at pt. 2, § 9(1).

[78] *Id*. at pt. 3, § 10, & pt. 4, § 20.

[79] *Id*. at pt. 4, § 21(2)(b).

[80] *Id*. at pt. 4 §§ 27–28.

[81] *Id*. at pt. 9 § 60.

4. Malaysia's Anti-Fake News Act

In April 2018, Malaysia enacted an "Anti-Fake News" law just ahead of national elections that resulted in a change in government.[82] The law prohibits the malicious creation, offering, publication, printing, distribution, circulation or dissemination of "fake news," which it defined as "any news, information, data and reports, which is or are wholly or partly false . . . ."[83] It defines these acts broadly such that they could include "re-tweets," forwards, and other actions that may not constitute endorsement.[84] Violators could face significant fines and up to six years in prison.[85]

The law also appears to create liability for intermediaries, including page administrators and moderators, that fail to immediately remove content "after knowing or having reasonable grounds to believe" that it contains fake news.[86] Finally, the law appears to allow for extraterritorial application where fake news "concerns Malaysia or the person affected by the commission of the offence is a Malaysian citizen."[87] The new, current government has attempted, so far unsuccessfully, to repeal the law.[88]

V. TAKING STOCK AND MOVING FORWARD

A. Corporate Approaches to Date

The corporate initiatives summarized above vary slightly at the margins but at their core are all primarily reactive, supply-side approaches. Given the "definitional" and "intent" challenges explained above, they have all wisely resisted censorship-oriented reactions, focusing instead on identifying and contextualizing potential disinformation. While this may help some users reconsider information they might otherwise believe, there is also evidence that for many users these approaches are at

[82] *See* Trinna Leong & Nadirah H. Rodzi, *Malaysia Passes Anti-Fake News Bill*, STRAITS TIMES (Apr. 2, 2018, 05:52 PM), https://www.straitstimes.com/asia/se-asia/malaysia-votes-in-anti-fake-news-law [https://perma.cc/KK75-AU75].

[83] ARTICLE 19, MALAYSIA: "ANTI-FAKE NEWS ACT" 9 (Apr. 2018), https://www.article19.org/wp-content/uploads/2018/04/2018.04.22-Malaysia-Fake-News-Legal-Analysis-FINAL-v3.pdf [https://perma.cc/F578-J2WQ] [hereinafter ARTICLE 19 LEGAL ANALYSIS] (quoting Section 2 of Malaysia Anti-Fake News Act).

[84] *See id.*

[85] *Id.* at 11–12.

[86] *Id.* at 12–13.

[87] *Id.* at 16 (citing Section 3 of the Act).

[88] *See* THE LAW LIBRARY OF CONG., INITIATIVES TO COUNTER FAKE NEWS IN SELECTED COUNTRIES 68–69 (Apr. 2019), https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf [https://perma.cc/QPS6-KJ9W].

best ineffective and at worst can reinforce the very content they are meant to discount.[89]

Sophisticated efforts to demonetize and de-platform repeat abusers have made the business of disinformation more difficult and should be enhanced. However, these efforts may be hampered by the platforms' lack of appetite to go after news sources that peddle disinformation but are seen as providing "ideological diversity."[90] In addition, given the "harm" challenge discussed above, the most persistent abusers are often able to easily regroup, tweak their tactics, and re-engage.[91] Unless and until there are significant improvements in attribution, these approaches are likely to continue to have limited impact.

More promising are the sophisticated approaches to detecting and neutralizing large-scale, coordinated disinformation campaigns (i.e., coordinated inauthentic behavior).[92] However, platforms still face serious challenges disaggregating harmful disinformation from legitimate speech, including satire and irony, across thousands of languages and dialects. Given this, and the tendency for disinformation campaigns to overlap with and reinforce existing authentic content, it is likely that these top-down, data-driven approaches will unintentionally lead to some degree of "over removal" and/or be manipulated by authorities to target legitimate speech, causing potentially serious freedom of expression consequences.[93]

As the European Commission has noted, efforts to empower consumers and researchers, including by providing detailed data about disinformation campaigns, have received far less priority to date.[94] When it comes to digital literacy, companies whose businesses hinge on understanding the tendencies and preferences of their users are well placed to produce effective, attractive campaigns, as well as to measure their effectiveness. However, to the extent certain prophylactic tactics may

[89] *See* Samuel Woolley & Katie Joseff, Demand for Deceit: How the Way We Think Drives Disinformation 11, 23 (Jan. 2020) (working paper) (on file with the National Endowment for Democracy), https://www.ned.org/wp-content/uploads/2020/01/Demand-for-Deceit.pdf [https://perma.cc/B558-9UY6].

[90] *See* Omer Benjakob, *Why Wikipedia Is Much More Effective than Facebook at Fighting Fake News*, HAARETZ (Jan. 9, 2020, 11:43 PM), https://www.haaretz.com/us-news/.premium-why-wikipedia-is-much-more-effective-than-facebook-at-fighting-fake-news-1.8378622 [https://perma.cc/4A98-BCGY] (discussing Wikipedia's decision to designate a number of pro-Trump news outlets as unreliable).

[91] *See, e.g.*, Warzel, *supra* note 32.

[92] *See supra* Section IV.A.36.

[93] *See, e.g.*, Layli Foroudi, *In Algeria, 'Electronic Flies' Threaten a Protest Movement*, .CODA (Dec. 10, 2019), https://codastory.com/disinformation/algeria-election-protest/ [https://perma.cc/2BPX-G39D] (describing how Facebook removed a "pro-democracy activist" who posted news and poetry).

[94] *See* FACEBOOK: CODE OF PRACTICE, *supra* note 47, at 51.

contradict those companies' financial interests, critics can be forgiven for being skeptical that they are the best messenger.[95]

Enhancing the ability of credible, objective third parties, such as independent academics, to examine the tactics and impact of disinformation campaigns, user behavior, and efforts to address the former by modifying the latter is vital. Although there have been some efforts in this direction, much more is needed. To the extent that companies cite data protection and other laws as barriers to such practices, they must be clearer and more explicit as to what precisely is impeding progress.[96] At the same time, policy makers should act quickly to clarify or rectify any such restrictions.

B. Government Approaches to Date

Government efforts to address disinformation have demonstrated a diversity of strategies. The EU's approach deserves praise for its open and consultative nature, as well as its inclination to resist resorting immediately to regulation (for now). This effort has had clear and tangible impacts on company behavior[97] and has yielded valuable transparency about the quality and quantity of disinformation campaigns on the major content platforms, as well as the efforts by the platforms to address them.[98] However, the Code has also been criticized on the one hand for allowing too

[95] *See* Rachel Kaser, *Facebook's Digital Literacy Library Is an Imperfect Course in Online Etiquette*, NEXT WEB (Aug. 2, 2018), https://thenextweb.com/facebook/2018/08/02/facebooks-digital-literacy-library-is-an-imperfect-course-in-online-etiquette/ [https://perma.cc/W3QP-Q8ZK].

[96] Early this year, the European Data Protection Supervisor issued an "Advisory Opinion," noting that "[t]here are concerns that the references to fundamental rights in the Code of Conduct on Disinformation could be a cover for . . . attempts to avoid scrutiny. . . . Resistance to greater transparency and accountability is justified [by social media companies] on questionable grounds of data protection," and it appears that "the reluctance to give access to genuine researchers is motivated no[t] so much by data protection concerns as by the absence of business incentive to invest effort in disclosing or being transparent about the volume and nature of data they control." EUROPEAN DATA PROT. SUPERVISOR, A PRELIMINARY OPINION ON DATA PROTECTION AND SCIENTIFIC RESEARCH 9 (Jan. 6, 2020), https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf [https://perma.cc/AX4E-EFLD].

[97] *See* Peter H. Chase, The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem 14 (Aug. 29, 2019) (working paper) (on file with the Transatlantic Working Group on Content Moderation Online and Freedom of Expression), https://www.ivir.nl/publicaties/download/EU_Code_Practice_Disinformation_Aug_2019.pdf [https://perma.cc/38KC-5WAG] (noting that the Code of Practice and associated efforts by the European Commission "are clearly making a difference in the behavior of the largest platforms on advertising, system integrity, public education and research access").

[98] *See generally* GOOGLE DISINFORMATION REPORT, *supra* note 37 (discussing Google's practices on disinformation).

much flexibility on the part of the platforms[99] and on the other for delegating too much authority and generally relying too much on platforms.[100]

The French government took a comparatively narrow approach by creating a time-limited (within three-months of elections) mechanism that keeps the adjudicatory function within the judicial branch, rather than outsourcing it to executive branch officials (as under POFMA), or relying on users to identify and platforms to adjudicate under the shadow of potential liability (like the notice-and-take-down approach in Malaysian law). However, this does not mean the law has been without controversy,[101] and it is too soon to know how the newly instituted "duty of cooperation" will work and impact freedom of expression.

Given the change in government and the lingering (although increasingly stale) promise to significantly revise if not repeal Malaysia's Anti-Fake News Act, it is not surprising that there is not much clarity as to its implementation. However, the breadth of the definitions, the vagueness of key aspects, the creation of an intermediary liability regime, and the harshness of the penalties in the law have all generated serious criticism.[102]

Singapore's POFMA has only been in force for a few months, but its early application has already generated considerable controversy.[103] To date, it appears that five of the six initial applications of the law have been against political figures or parties, which has confirmed the fears of many, notwithstanding the government's

[99] *See, e.g.*, SOUNDING BOARD OPINION, *supra* note 62.

[100] *See* ALEX KRASODOMSKI-JONES ET AL., DEMOS, WARRING SONGS: INFORMATION OPERATIONS IN THE DIGITAL AGE 10–11 (2019), https://demos.co.uk/wp-content/uploads/2019/05/Warring-Songs-final-1.pdf [https://perma.cc/38KC-5WAG]; Aleksandra Kuczerawy, *Fighting Online Disinformation: Did the EU Code of Practice Forget About Freedom of Expression?*, 6 DISINFORMATION AND DIGITAL MEDIA AS A CHALLENGE FOR DEMOCRACY: EUROPEAN INTEGRATION AND DEMOCRACY SER. (forthcoming 2020) (manuscript at 9–13), https://ssrn.com/abstract=3453732 [https://perma.cc/955X-7KB6].

[101] *See, e.g.*, David Kaye, U.N. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Letter dated May 28, 2018 from the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression addressed to the Government of France, U.N. Doc. OL FRA 5/2018 (May 28, 2018), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-FRA-5-2018.pdf [https://perma.cc/53B6-2W4D] (discussing France's law governing disinformation).

[102] *See* ARTICLE 19 LEGAL ANALYSIS, *supra* note 83, at 7–8; David Kaye, U.N. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Letter dated April 3, 2018 from the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression addressed to the Government of Malaysia, U.N. Doc. OL MYS 1/2018 (Apr. 3, 2018), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL_MYS_03.04.18.pdf [https://perma.cc/53B6-2W4D].

[103] *See* Kristen Han, *Want to Criticize Singapore? Expect a 'Correction Notice,'* N.Y. TIMES (Jan. 21, 2020), https://www.nytimes.com/2020/01/21/opinion/fake-news-law-singapore.html [https://perma.cc/U5FG-K5MW].

contention that this is simply a "coincidence."[104] The sixth and most recent application is against several news sources, as well as a Malaysian-based nonprofit that has refused to comply, potentially setting up a first blocking order under the law.[105] While it is always difficult to measure "chilling effects," given the generally weak protections for freedom of expression and media in Singapore,[106] it seems quite likely that the law and its early application are having significant impacts on freedom of expression.

Taken together, the limited set of government initiatives surveyed here suggests several observations. First, it is very difficult to appropriately address disinformation without risking serious negative implications for freedom of expression. As the five special mandate holders on freedom of expression and/or the media have noted, "the human right to impart information and ideas is not limited to 'correct' statements . . . [and] protects information and ideas that may shock, offend and disturb."[107] While these controversial ideas can be important in their own right, an even greater concern in democratic systems is the need to preserve the space to develop, express, and debate different ideas, which requires a degree of "intellectual privacy" that heavy-handed approaches to disinformation can inhibit.[108] This is especially true in countries that are not considered consolidated democracies. While narrower approaches, such as the one taken by France, may mitigate some of these potential consequences, they may also limit their intended impacts as well.

Second, efforts to work collaboratively with platforms through voluntary frameworks such as the "Code of Practice" can have impact but are limited both in scope (i.e., only those companies that "voluntarily" participate) and effect (companies will only go as far as they choose to go). In addition, it is unlikely that smaller political entities and/or markets will have the clout, expertise, and stamina to successfully engage in similar efforts.

Finally, although all of these initiatives have their respective faults (and some of them have many), they each share an important characteristic insofar as they address disinformation on its own, without grouping it together with other types of prohibited or problematic content. Given the unique characteristics that set disinformation apart from other categories of online content (the "definition," "intent," and "harm" challenges, discussed above in Section III), this approach

[104] John Geddie, *Coincidence that Fake News Law Applied to Politicians, Singapore Minister Says*, REUTERS (Jan. 6, 2020), https://news.yahoo.com/coincidence-fake-news-law-applied-082913894.html [https://perma.cc/8P3G-G9KL].

[105] *See* John Geddie, *Singapore Rebuts Illegal Hanging Report, Serves Fake News Notices*, REUTERS (Jan. 21, 2020, 10:48 PM), https://www.reuters.com/article/us-singapore-execution/singapore-rebuts-illegal-hanging-report-serves-fake-news-notices-idUSKBN1ZL0I3 [https://perma.cc/Y2SJ-BLLB].

[106] *See Freedom in the World 2019: Singapore*, FREEDOM HOUSE (2019), https://freedomhouse.org/country/singapore/freedom-world/2019 [https://perma.cc/7YR3-64N9] (last visited Apr. 13, 2020) (rating Singapore as "Partly Free").

[107] Joint Declaration, *supra* note 4, at 1.

[108] *See generally* Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387 (2008) (arguing for legal protection of records of intellectual activities under the First Amendment).

makes sense. While it is understandably tempting to develop one regulatory regime to rule all forms of online content, policy makers and legislators would be wise to consider disinformation on its own merits and learn from the experiences of those who have attempted to address it to date.