TRANSATLANTIC WORKING GROUP

The Santa Monica Session

Second session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, May 9-12, 2019, in Santa Monica, California

Co-Chairs Report No. 2

Key findings and recommendations
—*Susan Ness and Nico van Eijk*

Working Papers

The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem
—*Peter H. Chase*

Actors, Behaviors, Content: A Disinformation ABC
—*Camille François*

A Cycle of Censorship: The UK White Paper on Online Harms and the Dangers of Regulating Disinformation
—*Peter Pomerantsev*

Design Principles for Intermediary Liability Laws
—*Joris van Hoboken and Daphne Keller*

An Examination of the Algorithmic Accountability Act of 2019
—*Mark MacCarthy*

U.S. Initiatives to Counter Harmful Speech and Disinformation on Social Media
—*Adrian Shahbaz*

TRANSATLANTIC WORKING GROUP

Co-Chairs Report No. 2: The Santa Monica Session

Susan Ness, Annenberg Public Policy Center
Nico van Eijk, Institute for Information Law, University of Amsterdam

July 22, 2019

Introduction

The Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (TWG) convened its second session from May 9-12, 2019, at the Annenberg Community Beach House in Santa Monica, California.

Our first session, held February 27-March 3 at Ditchley Park in the United Kingdom, focused primarily on analyzing U.S. and European approaches to freedom of expression, and how these approaches could inform the ongoing initiatives to address hate speech and terrorism online. In particular, it examined the experience of four key initiatives to address online speech: Germany's Network Enforcement Act, or "NetzDG"; the EU's proposed Terrorism Content Regulation; the EC's Code of Conduct on Countering Illegal Hate Speech Online; and the Global Internet Forum to Combat Terrorism's "Hash-Sharing" Database. Our report on conclusions drawn from that discussion can be found here.

In Santa Monica, we reviewed recent developments in each of these areas, as well as the implications of the fallout from the tragic events in Christchurch, New Zealand. Among other things, we noted that:

- Increasingly, countries are moving toward statutory regulation of content moderation by online intermediaries, rather than improving the existing self- and co-regulatory mechanisms;
- Companies have tended toward over-removal (both based on their terms of service and in response to the increase in legally mandated short-removal times). They also lack independent oversight mechanisms for their content removal policies and practices under their terms of service and lack redress for such practices; and
- Current indicators of "success" for moderation policies, which tend to focus on the overall volume of content removed, are deficient. Other outcomes such as demonetizing content or reducing its visibility as well as the availability of redress should also be measured.

We then turned to the main theme of our second session: efforts to address "disinformation" or "viral deception," the term coined by Professor Kathleen Hall Jamieson to capture both intent to deceive *and* to disseminate. In contrast to illegal hate speech and incitement to violence, deceptive speech is not necessarily illegal. Arguably, it is protected in our transatlantic societies by freedom of expression and/or the First Amendment. That said, politicians, policymakers and the public increasingly see disinformation as causing serious societal harms, even when the content is not false but intentionally misleading. The rapid and broad (viral) dissemination may have been boosted artificially by "bots"

and fake accounts, by commercial actors (and sometimes even government officials), often with a malicious intent to weaponize social divisions, distrust in institutions, and other societal ills.

The group considered a number of specific initiatives that have either been adopted or are being considered in the United States and Europe to address disinformation:

- The EC Code of Practice on Disinformation;
- The United Kingdom's White Paper on Online Harms; and
- The Algorithmic Accountability Act, recently introduced in the Senate by Senators Ron Wyden and Cory Booker.

The TWG also discussed viral deception caused by or on behalf of a foreign government as part of an information operation. Government responses to information ops have a different set of tools available, ranging from diplomacy to sanctions to internet service denial.

Finally, the group began its review of intermediary liability in light of efforts underway in both Europe and the United States to condition or restrict the present forms of safe harbor for online platforms.

The background papers prepared for this session will be revised in light of the Santa Monica discussions and posted on the TWG website.

Key findings and recommendations

As co-chairs of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, we offer the following preliminary observations and recommendations, culled from the discussions in Santa Monica. There is no attribution, as the discussion proceeded under the Chatham House Rule. Members of the Working Group have reviewed our report and their comments generally are reflected in our conclusions below.

Adopt "Freedom of Expression by Design" as a guiding principle

First articulated at our Ditchley Park Session, this concept has even greater currency in the context of disinformation. Freedom of expression is a fundamental right underpinning our democracies, and is essential for holding governments accountable. Where speech – online or offline – clearly is illegal, it should be addressed according to applicable law.

When speech is not clearly illegal, governments must exercise extreme caution and refrain from requiring deletion either directly or indirectly (by essentially deputizing companies to take down offending speech). Governments and internet companies should consider positive measures instead, such as increasing both government and platform transparency regarding takedown requests, raising public awareness, investing in media literacy, and encouraging funding for high-quality news reporting and fact-checking. To avoid an actual or perceived conflict of interest, governments should refrain from directly funding news outlets or fact-checking organizations.

Focus on Actors and Behavior, rather than on Content that is odious but legal

The "A-B-C" analytical framework, which was presented in one of the papers and discussed during the meeting, helps policymakers to focus not on *Content* but rather on bad *Actors* and deceptive

Behavior. As content, "fake news" and "disinformation" are part of our democratic landscape, and are not *per se* illegal. Governments should not trigger the removal of undesirable, but legal, content, as such action is inconsistent with freedom of expression.

Governments may choose to address harmful behavior on the internet, where content is artificially propagated by bad actors – "bots, "astroturfers," or "troll farms" – as such conduct no longer reflects an authentic dialogue among citizens. This distortive behavior is akin to "spam," which companies have the technical expertise to address. But a cautionary note: some fake identities might be legitimate and should be protected, such as whistle-blowers calling attention to government corruption.

Strengthen enforcement of rules on foreign government interference

Foreign governments, too, have a right to have their voices heard in policy debates; that is diplomacy. But such engagement must be open and transparent. The United States as well as many European countries restrict interference from foreign governments in domestic political debates. Jurisdictions often prohibit foreign governments from making financial or in-kind contributions to political campaigns, and require foreign governments to register and report on their lobbying activities.

Covert foreign government manipulation of public opinion through artificial amplification and disinformation, or "information operations," is often deployed through multiple online channels and coordinated with real-world actions. These foreign governments may strive to deepen societal fissures by supporting both sides of contentious social issues. Such activities well may be illegal and better addressed through government channels, where additional tools are available, such as diplomatic pressure, sanctions, and disruption of internet service. Governments should determine whether and how to respond to such campaigns, bearing in mind that concerns about "information warfare" can be repurposed by authoritarian regimes to justify actions to impose "information sovereignty" within their borders.

Relevant European and American government agencies should strengthen collaboration against these "hybrid" tactics through NATO and other organizations, and should work cooperatively with companies and civil society to identify and derail such attacks.

Strengthen transparency and accountability

Companies should ensure that their terms of service and community standards are clear and accessible. Users whose content is deemed unacceptable and then removed or downgraded should be notified and provided a pathway for prompt redress. Both platforms and governments should disclose as much information as possible about enforcement actions taken.

In enforcing terms of service violations involving content, platforms should consider a variety of actions that lessen the impact on freedom of expression, including reducing content visibility through deceleration and demonetization, as well as deletion.

In an interim report in May, the French government suggested creation of a new regulatory regime to oversee both the transparency and accountability of platform content-moderation systems, rather than ruling on the content itself, to protect freedom of expression. It is an intriguing concept that deserves wider consideration.

During an election season, special attention should be given to both candidate and social issue advertising, as such communications are integral to the electoral process. If narrowly drafted,

governments could require specific disclosures for microtargeted candidate and social issue ads that state why the ad is being seen, the screening criteria, who paid for the ad, and the amount spent.

Transparency should require the logging and archiving of relevant data, to be made available for legitimate research purposes while guarding user privacy. Some platforms specifically block researchers from examining how their terms of service are enforced. Such restrictions should be lifted.

Consider an online court system or other independent body to adjudicate content moderation decisions

One proposal to resolve the sometimes conflicting roles of users and intermediaries is to create a system of specialized online courts that could quickly hear and adjudicate these disputes based on the digital record. These "e-courts" could be fast, simple and cheap; they would operate entirely online with no physical presence of complainant or defendant and no right of appeal (but still leave open the choice to file the case in the regular court system in lieu of the internet court). They would focus on whether content removal violated freedom of expression (based on the law of the complainant's jurisdiction); use specially trained magistrates; and, over time, build a public record of published decisions to serve as guideposts. Such a system could reduce the number of inappropriate removals, and could also protect platforms against undue government pressure to remove content that is troublesome but not illegal. The TWG will further develop this concept at its third session in November.

Separately, an independent body could be empanelled to review and redress cases of content removal or inappropriate termination of accounts and to provide guidance for platforms in novel situations such as the Christchurch attack. The selection of members, scope of authority, and scalability of social media councils are among the factors that the TWG should flesh out in the months ahead.

Be cautious if considering changing intermediary liability laws

Both in Europe under the e-Commerce Directive and in the United States under CDA Section 230, internet intermediaries have been protected to some degree against liability for content posted by users, in part to protect freedom of expression, but also to promote innovation and economic growth. These "safe harbor" protections are being revisited in Europe and North America, as legislatures and the public press for conditioning protection on proactive removal of troubling content. They want intermediaries to assume greater responsibility – a "duty of care" or even liability – for the actors, behavior and content on their platforms. The largest platforms often are viewed as controlling the public square.

The elimination of liability protections would likely result either in over-removal of lawful content, thus limiting freedom of expression, or passive posting of user content without moderation, thus elevating the amount of hate speech and viral deception online.

More nuanced approaches may offer alternatives to reducing intermediary liability protections. The TWG discussed an initial briefing paper on intermediary liability, which will be revised to participate in the public debate.

Promote media literacy, quality journalism, and fact checking

Viral deception is most effective when citizens are unaware of malicious attempts to influence their behavior. That impact can be reduced if the public knows how to identify stories that are false or misleading and promoted for malevolent ends. Governments have a duty to provide digital literacy education, not just for children but also for adults.

One tool in the fight against "disinformation" is serious fact-checking, although its scalability and effectiveness are limited. Major social media companies are investing in quality journalism and in respected fact-checking organizations. Platforms should be transparent about these efforts and protect the independence of these organizations. While elevation of trustworthy news sources is appropriate, there is a significant risk that lesser-known yet quality sources will be down-ranked, presenting risks to freedom of expression.

Governments should support and promote efforts to strengthen fact-checking organizations and journalism, provided that they maintain an arms-length relationship to preserve the independence of these entities.

Good Corporate Governance Encompasses Good Corporate Citizenry

Today's economy depends on a vibrant, global internet. Most internet companies, large and small, are legitimate and beneficial private-sector actors in our economies. To the extent that they give voice to the public by uploading their content, they contribute to democratic discourse and freedom of expression. But the internet has also spawned bad actors that take advantage of the openness of the network to rip apart the fabric of society.

As good corporate citizens, platforms should work proactively with policymakers and stakeholders to find scalable solutions to make the internet as safe and beneficial as possible while respecting freedom of expression. Solutions should take into account the size and variety of companies involved.

Governments should also strengthen consumer protection rules to ensure that platforms engage in appropriate behavior toward their users and other ecosystem companies.

Next Steps

Our final Transatlantic Working Group Session in November will examine:

- best practices in the use of artificial intelligence to address harmful content including algorithmic accountability;
- platform and government transparency;
- policy recommendations on intermediary liability; and
- policy recommendations for internet courts and social media standards councils.

During the third quarter, we will hold roundtables with stakeholders for additional feedback and engagement.

Our final report will be released at the end of the March 2020.

TRANSATLANTIC WORKING GROUP

The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem†

Peter H. Chase, Senior Fellow, The German Marshall Fund of the United States[1]

August 29, 2019

Contents

I. Summary

The EU Code of Practice on Disinformation is a government-initiated "self-regulatory" instrument that is unlikely to achieve its goal of curtailing "disinformation." The primary hurdle the EU (and other democratic societies) faces starts with the ambiguity surrounding the concept of disinformation, which makes it difficult to define the problem and devise appropriate counter-measures. For "disinformation" points to **content** deemed to have a pernicious effect on citizens and society even though that content is not itself illegal (unlike incitement to violence or child pornography, which are caught by other laws), and regulating it directly could undermine the fundamental right to freedom of expression. To skirt around this, the Code applies only to a small group of large platforms and advertisers' associations (not publishers or other parts of the information ecosystem); contains a

† One in a series: A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

limited series of measures that may curtail advertising revenue and the impetus that gives to the dissemination of certain content; and encourages transparency, system integrity, media literacy and research access. The Code does not, however, extend to the **actors** who create the content and drive disinformation campaigns, nor does it address the inauthentic **behavior** behind the rapid and widespread dissemination of that content – two critical elements that would help narrow the problem definition to the arguably more manageable issue of "**viral deception**."

While the efforts of the social media platforms pursuant to the Code had some impact in the run-up to the May European Parliament elections, disinformation, not surprisingly, is still seen as plaguing the public debate in Europe, leading to the likelihood of the next European Commission proposing more formal regulation of the social media platforms. But this too is likely to fail, as the Commission and European politicians are unlikely to find any level of disinformation acceptable. The result could all too easily be efforts to press platforms to take down more and more "harmful" – but not illegal – content, with all the implications for freedom of expression that implies.

The report that follows provides background and political context for the creation of the Code of Conduct in Section II; describes its main provisions in Section III; notes immediate reactions to and subsequent strengthening of the Code in Section IV; summarizes actions taken by the Code signatories in Section V; reports on the Commission's evaluation of the first half-year of the Code's operation as well as some of the other critiques in Section VI; adds some additional insights about disinformation in general in Section VII; and ends with some conclusions and recommendations in Section VIII.

II. Context and Background

The EU Code of Practice on Disinformation is a specific "self-regulatory" instrument to address the problem of disinformation in the European Union. But it is only the most recent manifestation of the EU's attempts to tackle the issue, and must be seen both in the context of that evolution, as well as part of a broader program to address it.

The European Union's fight against disinformation began with Russia's sustained attacks against Estonia in 2013.[2] When the then-new Commission led by Jean-Claude Juncker entered into office at the end of 2014, the overwhelming priority of generating growth and the renewed belief in the importance of European integration led to a distinction between the awareness that defenses against foreign actors were needed and the promotion of the "Digital Single Market" for economic reasons. The first moves against disinformation accordingly were with the creation of East StratCom in the EU's External Action Service (equivalent to a State Department/Ministry of Foreign Affairs) in March 2015, specifically to counter Russian disinformation narratives.

Reflecting this initial desire to separate (foreign) "fake news" campaigns from the internet as an economic instrument, the April 2016 speech by Commission Vice President Andrus Ansip (who hails from Estonia and led the Commission on digital policy) launching the Commission's Communication on Online Platforms only mentions "fake news" in connection with false advertising, or advertising counterfeit products.

This narrative evolved as the EU entered 2017, informed by the reports of Russian interference in the U.S. presidential elections. But in his remarks to the European Parliament, Ansip takes a decidedly measured tone even at that time:

> *Fake news – or simply "lies" – are also a serious problem. We are aware of the need to protect freedom of speech and to trust people's common sense. But we also need to be aware of the possible negative effects of this phenomenon…. Self-regulation and ethical standards play a very important role here. Social media platforms and users are acting to expose fake news and unmask the source. I also see global brands and media organisations deciding to move advertising money only to sites that are known to be free from harmful content. I welcome private sector initiatives to cut commercial funding of fake news sites….*
>
> *The concept of free speech protects not only that which we agree with – but also that which is critical or disturbing. We need to address the spread of false information by improving media literacy and critical thinking, as well as by better communicating why democracy, the rule of law, protection of minorities and fundamental rights are key interests for everyone. In all these actions, we have to bear in mind that it is our responsibility to protect fundamental rights, freedom of expression in the European Union. We have to believe in the common sense of our people. Once again, fake news is bad – but Ministry of Truth is even worse.*[3]

Two months later, by mid-June 2017,[4] the tone changed as the European Parliament in its resolution on the 2016 Communication on Online Platforms stressed the need to act against the dissemination of fake news; urged platforms to supply users with tools against it; and called on the Commission to analyze current EU law on fake news and to "verify the possibility of legislative intervention to limit the dissemination and spreading of fake content."

By the end of 2017, the Commission was moving in earnest on "fake news" as something far broader (and almost divorced from) the Russian threat, announcing on November 12 that it would establish a High Level Group on Fake News, and launching a public consultation the next day during a Multi-Stakeholder Conference on Fake News. In her speech to the conference, Commissioner Mariya Gabriel stressed that the internet brings many advantages, that the EU can't revert to the days of a centralized (and usually state-owned) media, and that educating consumers to identify fake news is critical. She also laid out four key objectives – transparency, diversity of information, credibility of information, and inclusiveness. The analysis of the nearly 3,000 responses to the consultation and the delivery of the report of the High Level Group in March led directly to the Commission's April 26, 2018, Communication – Tackling Online Disinformation: A European Approach, which in turn is the basis for the Code of Practice on Disinformation, adopted in September 2018.

Perhaps because it came out barely a month following the March 17, 2018, Guardian/New York Times splash about Cambridge Analytica, the Commission's Communication took a very different tone from the High Level Group report, which was published on March 12, five days before the story hit. The High Level Group importantly succeeded in shifting the narrative away from "fake news" to disinformation (indeed, the group's name was changed to include disinformation in the title), which it also defined as:

> *Disinformation … includes all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit.*

This definition has stuck, and many of the High Level Group's recommendations are reflected in the Commission Communication. But while the High Level Group Report emphasized the importance of protecting freedom of expression, warned against hard legislation to address a "multifaceted" and rapidly evolving problem, stressed the need for evidence-based decisions and even complimented social media platforms for the many steps they had already taken to address disinformation, the Communication argued:

> *(Social media) platforms have so far failed to act proportionately, falling short of the challenge posed by disinformation and the manipulative use of platforms' infrastructures. Some have taken limited initiatives to redress the spread of online disinformation, but only in a small number of countries and leaving out many users. Furthermore, there are serious doubts whether platforms are sufficiently protecting their users against unauthorised use of their personal data by third parties, as exemplified by the recent Facebook/Cambridge Analytica revelations, currently investigated by data protection authorities, about personal data mined from millions of EU social media users and exploited in electoral contexts.*

In stark contrast to the High Level Group report, the Communication also lashed out against the social media platforms for undermining the economic viability of traditional media (whereas the High Level Group stressed that traditional media, platforms and other actors should all be part of a broad "coalition" to address disinformation), noting *inter alia* that it will use reform of the EU copyright law to "ensure a fairer distribution of revenues between rights holders and platforms, helping in particular news media outlets and journalists monetize their content."

While the remainder of this report focuses on the Code of Practice on Disinformation, which the High Level Group recommended and which the Commission then pushed, it is important to note that this is just one element of the Commission's (and European Union's) approach to disinformation, which also includes a number of other specific measures in the five areas below:

Under a "**More Transparent, Trustworthy and Accountable Online Ecosystem**:"

- the Code of Practice (below);
- strengthening fact-checking by supporting the creation of an independent network of European fact-checkers based on the International Fact-Checking Network Code of Principles and by launching a secure European online platform to support their work;
- fostering online accountability through the EU regulation on electronic identification and uptake of IPv6, which allows the allocation of a single user per internet protocol address;
- harnessing new technologies, specifically artificial intelligence, to identify, verify and tag disinformation; tools to help citizens discover disinformation; technologies to help preserve the integrity of information; and cognitive algorithms to help improve the relevance and reliability of search results;

Under a "**More Secure and Resilient Election Process**:"

- this mainly involves working with the member states to ensure the integrity of their electoral infrastructure from cyberattack;

Under "**Fostering Education and Media Literacy**:"

- working with member states on media literacy programs;
- using the Audiovisual Media Services Directive mechanisms to monitor member states' engagement in this;
- expanding the EU's own programs on digital and media literacy;
- working with the OECD to add this as a criterion in its Program for International Study Assessments (PISA);

Under "**Support for Quality Journalism as an Essential Element of a Democratic Society**:"

- facilitate member state "horizontal" support (state aids) for quality media;
- provide additional EU-level funding for initiatives promoting media freedom and pluralism, quality news media and journalism;
- promoting a toolkit for journalists on ethical issues in addressing things like disinformation from a fundamental-rights angle;

Under "**Countering Internal and External Disinformation Threats Through Strategic Communication**:"

- provide additional resources to the EU External Action Service's East StratCom Task Force, the EU Hybrid Fusion Cell and the European Centre of Excellence for Countering Hybrid Threats;
- strengthen cooperation between these EU-level organizations and member states.

III. The EU Code of Practice on Disinformation

The Code of Practice on Disinformation ("the Code") was announced by the Commission on September 26, 2018, which heralded it as the first such (government-encouraged) self-regulatory initiative in the world.[5] The Code was the product of four months of deliberation among a working group of some of the larger online platforms and advertisers, with a "Sounding Board" including other stakeholders (media, civil society, fact-checkers and academia). Facebook (including Instagram), Google (including YouTube), Mozilla and Twitter as well as four key advertising associations participated in the exercise and were the initial signatories. Microsoft joined in May 2019.

In contrast to the May 2016 EU Code of Conduct on Countering Illegal Hate Speech Online, which establishes a clear set of obligations on all participants that was explicitly negotiated with the Commission, the Commission is not so obviously associated with the Code of Practice. The Code refers to and clearly takes guidance from statements in the Commission's April Communication, but it also quickly notes that (perhaps in contrast to the issue of illegal content) the signatories all work differently, and thus have different approaches to addressing content that is not illegal. As such, not all the obligations apply equally to all signatories.

Process: As noted above, the process which led to the development of the Code involved numerous opportunities for public engagement, including in response to the Commission's Communication, the formal three-month Consultation and Eurobarometer exercise, the conference and colloquia, the engagement of the companies and organizations subject to the Code, and the Sounding Board. That said, the Code itself was never presented to the broader public for comment before being published. And, as will be noted below, the Sounding Board participants unanimously rejected the Code as inadequate.

Scope: The Code defines disinformation as "verifiably false or misleading information, which, cumulatively, is created, presented and disseminated for economic gain or to intentionally deceive the public and may cause public harm, intended as threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security." Misleading advertising, reporting errors, satire or parody, and clearly identified partisan news and commentary are explicitly ruled out of scope.[6] The Code and its commitments apply only to the territories of the countries comprising the European Economic Area (the EU plus Norway, Iceland, and Lichtenstein).

Problem Definition: The only attempt at a problem definition in the Code is noting that the signatories agree with the Commission's conclusions that "the exposure of citizens to large scale Disinformation, including misleading or outright false information, is a major challenge for Europe. Our open democratic societies depend on public debates that allow well-informed citizens to express their will through free and fair political processes."

Evidence Base: No evidence is presented in the Code to establish that disinformation causes public harm, or that the platforms that are the focus of the Code "cause" such harm. To be fair, neither is any such evidence cited in the Commission's Communication.

The only attempts in the EU's discussion to provide such evidence are in the Synopsis analyzing the 2,986 comments received during the November 2017-February 2018 public consultation as well as a Eurobarometer poll of 26,576 residents in the 28 EU member states conducted in early February 2018 (so before the March 2018 Cambridge Analytica story hit). Both are essentially opinion- rather than evidence-based, and arguably neither supports the contention. While the Eurobarometer poll establishes that well over half the respondents believe they encounter fake news nearly every day (37%) or at least once a week (31%), and nearly 85% believe fake news presents a problem, three-quarters are totally or somewhat confident they can identify it (and so presumably are not swayed by it). Further, while large majorities in most member states trust traditional media, the percentage that "totally trusts" news from online social media hovers between 1% and 3% (5% maximum), while those who "tend to trust" it is about 26%. Otherwise, European citizens take a very skeptical eye to what they see on social media. (Interestingly, there seems almost to be an inverse correlation between the two sources of news: in countries where traditional media – often state-controlled – is not trusted, social media sources are.)[7]

The Synopsis of comments received goes into more depth. It echoes the poll in terms of sources of trust in different types of media (lots for traditional sources, less for social media), perceived exposure to fake news, and a strong ability to discern it. But the more detailed questionnaire and the opportunity for open responses from the 2,784 individuals and 202 legal entities (including 69 from news media,

51 from civil society and 16 from platforms) and journalists leads the Synopsis to a more serious definition of "fake news" and the problems it causes:

> *…based on the pursued objectives of the news. The concept would mainly cover the online news, although sometimes disseminated in traditional media too, intentionally created and distributed to mislead the readers and influence their thoughts and behaviour. Fake news would seek to polarise public opinion, opinion leaders and media by creating doubts about verifiable facts, eventually jeopardising the free and democratic opinion-forming process and undermining trust in democratic processes. Gaining political or other kinds of influence, or money through online advertising (clickbait), or causing damage to an entity or a person can also be the main purpose of fake news.*

Despite the malicious nature of fake news, and the drivers perceived behind its widespread dissemination (especially on social media), the Synopsis offers no evidence of its *actual* impact (although some of the written submissions may have cited research on this); rather it states:

> *Some civil society organisations noted that both the spread and the impact of disinformation are smaller than generally assumed and that more studies are needed to properly understand the phenomenon.*

Actors: The Code applies only to the signatories – Facebook (including Instagram), Google (and YouTube), Mozilla and Twitter as well as the advertising associations that signed, specifically the European Association of Communications Agencies (EACA), the Interactive Advertising Bureau of Europe (AIB) and the World Federation of Advertisers (WFA). The latter do not enter into obligations on behalf of their members, but undertake to educate them on the Code, and to encourage them to adhere to its principles.

Objectives: "In line with the Commission Communication," the signatories agree to 11 objectives, namely, to:

- include safeguards against disinformation;
- enhance scrutiny of advertisement placements to reduce revenues to purveyors of disinformation;
- ensure transparency of political and issue-based advertising and give users means to understand why they've been targeted for it;
- implement policies against misrepresentation;
- close fake accounts and mark bots' activities to ensure they're not mistaken for human activity;
- ensure the integrity of services against accounts that spread disinformation;
- prioritize relevant, authentic and accurate information;
- ensure transparency through indicators of trustworthiness of content sources, media ownership and verified identity;
- dilute the visibility of disinformation by improving the findability of trustworthy content;

- empower users to customize newsfeeds to facilitate exposure to different views and report disinformation; and
- facilitate access to data for research.

The above is a simplified version of the actual objectives, which are more nuanced (and less onerous) than noted here. These objectives track fairly closely to the 10 principles the High Level Group recommended.

Measures: The Code underscores that, given their differences, not all signatories can work to achieve each of these objectives. As such, the signatories variously "commit" to the extent they can to adopt 15 specific measures in five categories of action:

Scrutiny of Ad Placements

- Policies and process to disrupt advertising and monetization for disinformation activities;

Political and Issue-Based Advertising

- Ensure advertised content is presented as such;
- Public disclosure of political advertising (for a candidate or on a referendum), including sponsor and amount spent;
- Public disclosure of issue-based advertising (which needs a better definition);

Integrity of Services

- Adopt clear policies regarding identity and misuse of bots;
- Adopt policies on impermissible use of bots;

Empowering Consumers

- Invest in products, technologies and programs to provide effective indicators of trustworthiness;
- Invest in technological means to prioritize trustworthy content in search, feeds or other automatically ranked distribution;
- Invest in features and tools that allow consumers to find different perspectives;
- Partner to enhance digital and media literacy;
- Help market tools to help consumers understand why they're targeted by advertising;

Empowering the Research Community

- Support independent efforts to track disinformation, including by sharing data sets and undertaking joint research;

- Don't prohibit or discourage research into disinformation and political advertising on their platforms;
- Encourage research into disinformation and political advertising;
- Convene annual meetings of stakeholders, fact-finders and researchers into these issues.

Many of the specific steps the signatories currently take in each of these areas are spelled out in an Annex of Best Practices, which provides links and details to the numerous initiatives the companies and associations had undertaken.

Remedies/Mitigation: The Code has no provisions for remedial action against unjustified takedowns of content, but neither is this a point in the Commission's Communication.

Oversight: The signatories commit to meet regularly to assess developments under the Code, to provide an annual report on activities related to the measures above, and to evaluate the effectiveness of the Code after a year, when they will discuss continuation of the Code and possible follow-up. They commit as well to engage an "objective" third party to review their self-assessments and evaluate progress toward meeting the objectives in the Code. They also commit to cooperate with the Commission, including by providing information upon request, informing the Commission of new signatories or withdrawals, responding to questions and inviting the Commission to their meetings.

IV. Immediate Reactions and Subsequent Strengthening of the Code

As noted previously, the Sounding Board[8] that was consulted as the Code was developed unanimously believed it insufficient as it contained "no common approach, no clear and meaningful commitments, no measurable objectives or KPIs (key progress indicators), hence no possibility to monitor process, and no compliance or enforcement tool."

For its part, while the Commission welcomed and indeed heralded the Code of Practice when it came out in September 2018, by December it obviously had doubts. In a December 5 Report to the European Parliament and Council on the implementation of its April Communication, the Commission notes that the Code provides an "appropriate framework" for pursuing its objectives, and disagrees with the Sounding Board by saying that the Code is consistent with Commission principles for self-regulation.

Metrics: But it also used that report (as well as an accompanying Action Plan) to effectively order the signatories to report by the end of December on actions taken, and then to report monthly through the May 2019 European elections. Furthermore, it responded to the critique of the Sounding Board by spelling out Key Progress Indicators for each of the 15 commitments (number of accounts removed for violating advertising policies, number of websites blocked for scraping content, number of political ads taken down for failing to be transparent, number of records provided to repository, number of identified fake accounts, etc.). Because the Commission would use these KPIs to measure the "success" of the Code, they are replicated in full in Appendix I.

Oversight: It also announced that it will enlist the network of member state regulators responsible for overseeing implementation of the Audio-Visual Media Services Directive (as well as the European

Audio-Visual Observatory) to assist it in monitoring compliance. Finally, the Commission signaled that "(s)hould the results prove unsatisfactory, the Commission may propose further actions, including of a regulatory nature."

V. Platform Actions Under the Code

The monthly reports the three major social media platforms that are signatories to the Code (Facebook, including Instagram; Google, including YouTube; Twitter) were compelled to issue under the Commission's Action Plan provide insight into the actions taken (in part in response to the Code but also for the companies' broader interests), and the impact those actions may have had with respect to the right to freedom of expression.

And clearly the Code had an impact. The monthly reports[9] of Facebook, Google and Twitter are structured to describe the efforts they instituted to address each of the five groups of measures that the Commission was focused on – scrutiny of ad placements; political and issue-based advertising transparency; service integrity; empowering consumers; and empowering research communities. A more detailed summary of the steps the three companies took in each of these areas is provided in Appendix 2, but some of the highlights include:

- "Google took action against 131,621 EU-based ads accounts for violating its misrepresentation policies, and against 26,824 EU-based ads accounts for violating its policies on insufficient original content; it also took action against 1,188 EU-based publisher accounts for violating its policies on valuable inventory. Facebook reported on some 1.2 million ads actioned in the EU for violating its policies on low quality or disruptive content, misleading or false content, or circumvention of its systems. Twitter reported rejecting 6,018 ads targeted at the EU for violation of its unacceptable business practices ads policy as well as 9,508 EU-targeted ads for violations of its quality ads policy;"[10]

- All three companies in March instituted new procedures to "verify" that those who want to place political ads are legitimate European-based individuals/entities, and all three by May had developed online searchable databases for these ads for all EU member states;

- Facebook also instituted similar requirements for issue-based ads related to immigration, political values, civil and social rights, security and foreign policy and environmental politics, all of which are again searchable in its Ad Library database;

- On integrity of services, Google reported taking down literally millions of YouTube channels for violating its misrepresentation and impersonation policies, while Facebook described in detail its efforts against "Coordinated Inauthentic Behavior" (including under a number of Russian-based campaigns) and noted that it took down 2.19 billion fake accounts (worldwide) during the first quarter of 2019, and Twitter reported challenging 76.6 million spam/bot/fake accounts and acting on another 2.3 million accounts reported by its users in the first five months of 2019 (again, worldwide);

- In terms of "empowering consumers," all three companies report significant efforts to support the fact-checking community, promote quality news (demote low-quality content), educate

thousands of journalists, conduct digital-media literacy campaigns for nearly a million European citizens, and help European politicians and political campaigns to protect their websites and digital communications from malicious actors;

- Finally, the companies cite efforts to help the research community, including granting access to their ads transparency websites, as well as numerous specific research projects they have funded related to disinformation, in Europe and elsewhere.

VI. Evaluating the Code

The Commission's initial evaluations of the early monthly reports submitted by Facebook, Google and Twitter were stinging. The Commission in February said it "remains deeply concerned by the failure of the platforms to identify metrics that would enable the tracking and measurement of progress in the EU as well as by lack of sufficient detail on the platforms' plans to ensure that actions in pursuit of their policies are being deployed in timely fashion and with appropriate resources across all Member States." It repeatedly "regrets" (sometimes even "deeply") that the companies did not supply sufficient information or metrics. And while in March it "takes note" of the progress the platforms made in their second monthly reports (especially on political ad transparency tools), it also stresses that "further efforts are needed by all signatories," especially in providing further metrics and details.

In its June Intermediate Targeted Monitoring evaluation of the reports the companies had submitted through the May EP elections, the Commission more positively welcomes the work of the companies, and notes:

> *The policies on transparency for online political ads implemented by the platforms as well as their actions against malicious bots, fake accounts and coordinated inauthentic behavior have likely helped limit the impact of disinformation operations from foreign and domestic actors. This is supported by a number of studies and independent sources, which suggest that the dissemination of disinformation in the run up to the European elections was not alarmingly high. For instance, according to a study by the Oxford Internet Institute, which carried out a thematic analysis of the top 20 junk news stories on Facebook and Twitter, fewer than 4% of news sources shared on Twitter ahead of the 2019 EU elections was junk news, while mainstream professional news outlets received 34% of shares. According to FactCheckEU, the European branch of IFCN, there was less disinformation than expected in the run up to the European elections and it did not dominate the conversation as it did around the past elections in Brazil, the UK, France or the United States.*

In its more formal June 14 Communication to the European Parliament and Council, the Commission reaffirms that the Code of Practice and other aspects of the Action Plan "contributed to deter attacks and expose disinformation…. raising awareness about how to counter the threat. Increased public awareness made it harder for malicious actors to manipulate the public debate." Yet the Commission acknowledges, in citing a report from Avaaz and the Institute for Strategic Dialogue, that these efforts did not stem the disinformation tide:

> *More than 600 groups and Facebook pages operating across France, Germany, Italy, the United Kingdom, Poland and Spain were reported to have spread disinformation and hate speech or have used false profiles to artificially boost the content of parties or sites they supported. These pages generated 763 million user views.*

Reports from researchers, fact-checkers and civil society also identified additional instances of large-scale attempts to manipulate voting behaviour across at least nine Member States.[11]

The Commission continues to press the platforms for additional ad transparency (particularly Google and Twitter on issue ads), more collaboration with fact-checkers and news trustworthiness indicators, and more cooperation with researchers. And again it stresses that "[s]hould the results of this assessment not be satisfactory, the Commission may propose further initiatives, including of a regulatory nature."

The Commission does not comment in its evaluations of the Code on the impact it might have had on freedom of expression. But the companies' own reports about the implementation of their "verification" process of potential political advertisers should have raised questions. By the end of May, Twitter had certified only 27 political advertising accounts; of the 676 applications Google had received, only 174 had been verified (and had run some 75,000 ads, generating €3.9 million in revenues). These numbers imply quite a lot of speech that did not benefit from amplification during the elections process, and both companies acknowledge that many of these applications were likely from legitimate sources whose applications were denied pending additional documentation. In another area, Facebook (which does not report how many political advertising accounts in Europe it did not verify) is now publishing the results of appeals of content removals that it later determined were unjustified: against the 1.1 million pieces of content removed worldwide during the first quarter of 2019 for violating Facebook's hate speech community standards, over 152,000 pieces, or 13.9 percent, were subsequently reinstated.[12] While the company's transparency on its appeals, review and reinstatement record is laudable, a more than 10 percent wrongful removal rate represents a not-insignificant impact on freedom of expression, including in the European Union.[13]

Other Commentary: Much of the other commentary about the Code published since the companies' baseline reports has criticized its voluntary nature and self-regulation in general, while stressing the importance of the long-term media-literacy efforts. A number of different sources also complain that the companies' attempts at greater transparency are still insufficient, including with respect to their political ad transparency efforts.[14] One of the more thoughtful pieces, by Paul Butcher of the European Policy Center, is more positive about the self-regulatory efforts in part as government regulation can be even more ham-fisted. He argues for greater publicity of the Code and its reports so the potential of broader public criticism (and its effect on share prices) can help hold the platforms to account, and recommends the platforms in general to be more forthcoming to such public oversight, including by civil society and researchers.

VII. Other Analyses

The EU Code of Practice on Disinformation of course was developed in the context of a much wider global debate about disinformation and its impact on broader society that goes well beyond the scope of this paper. Much of that commentary does not talk explicitly about the Code of Practice, although the analysis in it of course is pertinent to an understanding of the Code.

For instance, a recent study by the Center for the Analysis of Social Media in the British think tank Demos, Warring Songs: Information Operations in the Digital Age, provides an analysis of 39 case

studies of systemic disinformation efforts across 19 countries, including in-depth reports on those cases in three European countries.[15] The report underscores that much of the content shared in disinformation campaigns is not "fake," but selective amplification of reputable, mainstream media stories to fit an agenda. As such, it argues, the focus on fake news and disinformation is "myopic," as "information operations are vast in scale, varied in target and numerous in strategies and tactics." It goes on to define the problem more precisely as:

> *A non-kinetic, coordinated attempt to inauthentically manipulate an information environment in a systemic/strategic way, using means which are coordinated, covert and inauthentic in order to achieve political or social objectives.*

The Demos report further provides a taxonomy of the aims, strategies and tactics of such operations:

Aims, Strategies and Tactics of information operations

Aims	Strategies	Tactics
Affect sympathetic changes in behaviour and perception	Build political support Feign public support Encourage conspiratorial thinking Promote sympathetic voices	Astroturfing (fake grassroots support) False amplification of critiques of opponents False amplification of marginal voices False amplification of news Impersonation of public figures Impersonation of political allies
Reduce oppositional participation	Reduce critical voices in media Undermine trust in political representatives and institutions Undermine trust in institutions of government Undermine trust in electoral institutions Incite societal and cultural divisions Voter suppression Abuse of legal systems	Defamation Doxxing Hacking and leaking documents Interference with political processes Intimidation and harassment Dark advertising
Reduce quality of communications environment	Create confusion and anger Denigrate compromise Undermine channels of productive communication Reduce trust in digital communications Disrupt channels of communication	Exploitation of content moderation systems Playing both sides Scare stories Shocking or graphic content Communications disruption Hashtag poisoning Spam
Reduce quality of available information	Undermine trust in media institutions Undermine trust in digital media Blur the boundaries of fact and fiction Suppress critical content Promote sympathetic content Shift the balance of content in actor's favour	Algorithm exploitation and manipulations Deepfakes Dissemation of doctored images, videos and documents Dissemation of false, misleading or misattributed content Impersonation of websites Restriction of availability of information to the public Dissemation of conspiracy theories

A similar analysis of the breadth of the issue in the context of the European elections (albeit focused more on Russia as a malicious actor) highlights that "the tools and channels used to deliver disinformation to an audience will be different, and social media is not always the most important channel …. Social media platforms do not produce the malicious content; they just are used and abused to spread it. Social media may be a very powerful weapon, but the platforms are not the ones pulling the trigger."[16]

This points to an extent to a problem exacerbated by the range of actors included in the Code. For as important as the large platforms are to Europe's social discourse, they do not have monopolies –

indeed, smaller platforms can have as much (if not more) impact on sub-national/linguistic/regional polities, as can many other sources of news, including traditional media.[17]

VIII. Conclusions and Recommendations

The Code of Practice on Disinformation is a fairly messy and in some ways structurally incoherent document, but the strong political pressure behind it and the Commission's efforts to strengthen it in December 2018 with stricter reporting requirements and more oversight are clearly making a difference in the behavior of the largest platforms on advertising, system integrity, public education and research access.

While all these efforts will cut into the profits of the large platforms by reducing some advertising revenues and increasing compliance costs, they will not solve the "problem," which was poorly formulated and not well substantiated.

Disinformation by its very nature is content, in this case defined as having malicious intent – where intent is difficult to discern (although that intent can be imputed, and often is if the content isn't liked). And just as it is difficult to regulate content that is not illegal (and the Code explicitly acknowledges disinformation is not illegal), regulating only how large platforms disseminate some types of content (essentially, constraints on monetization of certain ads) will not be effective. It cannot and will not capture all malicious content (never mind "undesirable" content, which is what many politicians are concerned about); as such, it can't prevent all – or perhaps even most – of the worst instances of "viral deception."

As such, while the Code helped demonstrate the Commission was "doing something" in the run-up to the EP elections, and nudged the large platforms into better practices in a number of laudable respects, it is highly likely to be judged wanting. Indeed, the new European Commission that will enter office November 1, 2019, under the newly nominated President, Ursula von der Leyen, will propose new "hard" legislation (a "Digital Services Act") of social media platforms in part to address the problem of disinformation. This is likely to include a revision to the EU's e-Commerce Directive, which, like Section 230 under U.S. law, exempts online intermediaries from liability,[18] thus increasing compliance costs (including on small platforms that may not have the resources to make necessary technical changes, thereby increasing large platform dominance).

*****

European – and American – politicians and policy-makers are right to be concerned about the deep divisions that are appearing in their societies. They are correct as well in understanding that the rapidity with which messages spread in the online environment can exacerbate these divisions.

But they need to think carefully about their policy prescriptions to address these "harms," especially when looking at disinformation. They may not like or agree with certain content, but should bear in mind former Vice President Ansip's admonition, quoted above, that "Fake news is bad – but a Ministry of Truth is even worse."

First and most importantly, they need to distinguish between the message and its reception. If a piece of content resonates with a section of the public, whether or not that message is "factual," they need

to ask and understand why. This will not be easy, as it may point to deeper societal problems that our political systems find difficult to address. But in many ways those are the real problems, whether they are mistrust of the elites, doubts about the effectiveness of institutions (including the European Union or Congress), migration and fear of foreigners (the main themes seen in most reports about the European elections), or something else.

Second, they need to differentiate between pieces of content ("disinformation") and disruptive campaigns, that is, information operations that use the internet to generate viral deception. Here identifying the right actor is critical. Social media platforms of all sizes may be vectors for (parts of) these campaigns, but they are not the villains. Rather, those behind the campaigns are, whether they are foreign or domestic government or non-state actors using false information or selective presentation of true. Addressing the activities of those actors directly, through such efforts as the EU's StratCom Task Force or more powerful legislative acts, is arguably more important. The platforms are allies in this fight, as they need the trust of their communities and clients (advertisers) to succeed, and have the tools to disrupt at least some of the inauthentic behavior/amplification behind the campaigns. (In that sense, politicians should be as concerned about ham-handed over-removals as they are of insufficient action.)

And, as the EU's High Level Expert Group emphasized, platforms are only one part of the broader internet ecosystem that needs to be enlisted in this effort. Concerns about social media platforms taking advertising revenue from traditional media have no place in the disinformation discussion, however valid worries about the commercial health of traditional media might be.

Finally, politicians may need to admit to their publics that the "disinformation problem" cannot be resolved through self-regulatory, co-regulatory or even legislative means. This does not mean giving up. The longer-term efforts of governments, platforms and other parts of society to build media and digital literacy and to support additional research on the nature of information operations and viral deception are critical. But citizens in the end need to know that they are their own best defense, and accept that responsibility – if they are to protect their fundamental right to freedom of expression.

Appendix 1: Commission Key Performance Indicators for Code Signatories

A. Scrutiny of ad placements	
1. Deploy policies and processes to disrupt advertising and monetisation incentives for relevant behaviours	• Number of accounts removed for violation of platform advertising policies (e.g. policies against misrepresentation) • Policies put in place to demote sites or accounts that distribute disinformation or inauthentic information (e.g., click-bait) • Percentage of contracts between advertisers and ad network operators with brand safety stipulations against placement of ads on disinformation websites • Number of websites blocked for duplicating or "scraping" content produced by other websites
B. Political advertising and issue-based advertising	
2. All advertisements should be clearly distinguishable from editorial content	• Ads properly labelled as political advertising as a % of overall political ads • Actions taken to ensure all political ads are properly labelled • Number of political or issue-based ads taken down for failure to comply with platform guidelines on the transparency of political advertising
3. Enable public disclosure of political advertising	• Number of records added to public disclosure repositories • Information on amounts received from political parties, candidates, campaigns and foundations for political or issue-based advertising • Policies to verify the identity of political ads providers
4. Devising approaches to publicly disclose "issue-based advertising"	• Information on progress on this commitment
C. Integrity of services	
5. Put in place clear policies regarding identity and the misuse of automated bots on their services	• Number of identified active fake accounts • Number of identified active fake accounts disabled for violation of platform policies • Information on measures to ensure all bots are clearly labelled as such. • Number of posts, images, videos or comments acted against for violation

	of platform policies on the misuse of automated bots
6. Put in place policies on what constitutes impermissible use of automated systems	• Information on policies about the misuse of bots, including information about such bot-driven interactions • Number of bots disabled for malicious activities violating the platforms' policies

D. Empowering consumers	
7. Invest in products, technologies and programs […] to help people make informed decisions when they encounter online news that may be false	• Information on investments made in such tools or other progress towards this commitment • Information on actual use of such tools by consumers • Information on collaborations with media organisations and fact-checkers to carry out this commitment, including development of indicators of trustworthiness • Information on measures to make fact-checked content more visible and widespread.
8. Invest in technological means to prioritise relevant, authentic and authoritative information where appropriate in search, feeds, or other automatically ranked distribution channels.	• Information on progress on this commitment • Information on collaborations with media organisations and fact-checkers to carry out this commitment , including the development of indicators of trustworthiness
9. Invest in features and tools that make it easier for people to find diverse perspectives about topics of public interest	• Information on investments made in such tools or other progress towards this commitment • Information on availability of such tools and use of such tools by consumers
10. Partner with civil society, governments, educational institutions, and other stakeholders to support efforts aimed at improving critical thinking and digital media literacy	• Information about initiatives carried out or planned by signatories, including degree of coverage across Member States
11. Encourage market uptake of tools that help consumers understand why they are seeing particular advertisements	• Information on actual uptake of such tools and use by consumers

E. Empowering the research community	
12. Support good faith independent efforts to track Disinformation and understand its impact	• Information on collaborations with fact-checkers and researchers, including records shared
13. Not to prohibit or discourage good faith research into Disinformation and political advertising on their platforms	• Information on policies implementing this commitment
14. Encourage research into Disinformation and political advertising	• Information on policies implementing this commitment
15. Convene an annual event to foster discussions within academia, the fact-checking community and members of the value chain	• Report on the annual event

Appendix 2: Social Platform Actions Under the Code of Conduct

The Facebook, Google and Twitter "baseline" reports published in January 2019 are structured to allow the companies to provide more in-depth narratives about the efforts they instituted to address each of the five groups of measures noted above (ad placements, political and issue-based advertising transparency, service integrity, empowering consumers and empowering research communities), globally and within the European Union (including where certain activities are available only in some member states).

Facebook, for example, noted in its baseline report that when fact-checkers rate a story as false, it significantly reduces that story's distribution in News Feed, cutting future views by more than 80 percent. It also reported that it took down 98.5 percent more fake accounts in the second and third quarters of 2018 than in the first, 99.6 percent of which it had flagged itself (although most of these were related to commercially motivated spam). It further reported that in Belgium, it took down 37 pages and 9 accounts around the time of the local elections, some of which were initially identified by Belgian media as potentially inauthentic and trying to manipulate political discourse, as subsequent investigation further confirmed. Prior to the French presidential election in 2017, it removed more than 30,000 fake accounts that were engaging in coordinated inauthentic behavior to spread spam, misinformation or other deceptive content. Facebook also used its initial report to spell out in detail its efforts to address "Coordinated Inauthentic Behavior" (CIB), essentially content-agnostic actions to prevent the spread of content through bots and other means. **Google**, for its part, reported that in 2017 it had disapproved some 3.2 billion ads, blocked 2 million pages, terminated 320,000 publishers, and blacklisted 90,000 websites for overall content policy violations, including some 650 websites for violating its "misrepresentative content" policy. **Twitter** argued that it had made significant efforts to curb malicious automation and abuse, suspending more than 1,432,000 applications in 2018 (including 75 percent of the accounts challenged during the first half of the year). The **Mozilla** submission is shorter and specifically addresses its commitment to increase staff, roll out enhanced security features for its Firefox browser, support researchers, and launch an EP "Elections Bundle" to provide more transparency around political ad targeting. The **advertising agency** reports are primarily statements regarding efforts they have undertaken to publicize the Code among members.

In the subsequent monthly reports for January to May,[19] Facebook, Google and Twitter (the only subsequent reporters) clearly went much further. In all cases, the three platforms report on stepped-up actions against advertisers that don't meet their criteria, new processes for political (and in the case of Facebook, issue-based) ads, their focus on inauthentic behavior, and their work with politicians and the broader fact-checking and research community to find ways to detect and demote disinformation. In keeping with the Commission's emphasis in the December Report and Action Plan, the companies provide as many hard numbers as they can.

Advertising[20]

Facebook: Unlike Google and Twitter, the Facebook reports provide little statistical detail about advertisement removals on its social media platform and Instagram, beyond noting that in both March and April it "identified and actioned" over 600,000 advertisements to EU audiences that did not meet the company's standards on quality, content and/or procedures (including such things as "click-

baiting" with overly emotive images, deceptive promotion, etc.). The company argues in part that its policies and practice of reviewing ads before they are published prevents questionable ads from being shown.

More significantly, however, the company in late March launched its online Ad Library, which provides a searchable database of all ads being run on its platforms in selected countries – including, significantly, all 28 EU member states.

Google: Each of the monthly Google reports provides details (including by member state) of the number of EU-based ads and website publishers taken down for violating the company's policies on misrepresentation and content:

Issue/removals by month	January	February	March	April	May (to May 26)
Misrepresentation on Google Ads	48,642	20,627	10,234	35,428	16,690
Websites violating AdSense misrepresentation policies	0	1	0	2	0
Ads with problematic content	3,258	5,501	5,904	6,696	5,465
AdSense publishers with problematic content	205	215	370	310	88

Twitter: In the first three months, the Twitter report mainly summarizes its advertising policies; it only provides statistical data as of the April report. There, it reports that 4,590 ads that did not meet its unacceptable business-practices ad policy were prevented from being shown to European audiences during the first three months of the year, while 7,533 ads were blocked for not meeting the company's quality ads standards. The May report provides no details about April, but only about the first 20 days of the month, where the numbers are 1,428 and 1,975 respectively.

Political and Issue-Based Advertising

Facebook: Facebook in March also began its EU Political Ads process and transparency reporting, requiring verification of advertisers, labeling of ads, and transparency about their viewership. As of the end of May, there had been 343,726 political and issue ads promoted by Facebook across the EU, generating some €19 million in revenue for the company across its platforms; details for each of these ads (including about the demographics of those who viewed it) can be found on the Ad Library Report page for each of the EU member states, as well as Canada, India, Israel, Ukraine and the United States.

Google: The company brought the political ads verification process and transparency reports it had used in the United States to Europe in January 2019; the guidelines related to verifying the validity of political parties and candidates wanting to purchase ads were published in February. Subsequent reports to the Commission spell out how many applications there were to be a valid political advertiser, how many were verified/being reviewed/rejected (mainly for lack of appropriate documentation), and how many ads were approved, shown (not all approved ads by political advertisers were actually published) and rejected. The procedure for applications opened March 14, with the first labeled ads published as of March 21. As of May 28, 2019, the Google transparency report indicated that some 74,828 political ads had been shown to the European public on Google's various platforms (including YouTube) between March 21 and end May, generating €3.9 million in revenue for the company.

Google: Political Advertiser/Advertisement Data for Europe, by Month, 2019

Issue/Month	March	April	May
Advertiser applications received	120	556	676
Advertisers verified	18	123	174
Applications under review	16	13	57
Applications rejected	86	420	445
Ads approved	11,000	56,968	98,000
Ads Shown	--	10,289	63,000
Ads rejected	12,000	16,195	50,000

Twitter: Twitter's Political Advertiser Certification process and Political Ads Transparency Center began operating in Europe in March. In its May report, it notes it had received 66 applications in 11 EU member states to be certified to do political advertising in the EU; of these, 27 had been registered as of May 20. It also notes that 515 political ads (those mentioning parties or candidates in the European elections) were prevented from being shown to Europeans between March 11 and May 20 (12 of these are reported in the April report up until April 11).

Integrity of Service

Facebook: Where Twitter in many ways focused on its Political Ads Transparency initiative, Facebook's monthly submissions concentrate on its efforts to ensure the "integrity of services." This starts in part through the identification of fake accounts; in its March report, Facebook published that it had taken down 2.19 billion inauthentic accounts globally during the first quarter. It is also related to Facebook's work demoting visibility of stories that have been critiqued by fact-checkers, discussed below. More interesting is the extensive explanations Facebook provides across the reports on its work against "coordinated inauthentic behavior" on its platforms, much of which is associated with fake accounts; each month describes two to six different networks of disinformation operations that it disabled, including in Belgium, France, Kosovo, Macedonia, Moldova, Romania, Ukraine, and the United Kingdom. A substantial number of these were efforts linked to Russia, but certainly not all – in the case of Romania, for instance, the disabled networks included mainly fictitious accounts operating in support of the Social Democratic Party. Israel and Iran are identified as sources of inauthentic behavior in the May report. In this connection, Facebook announced stepping up penalties against those who abuse its CIB guidelines. Facebook also describes its work against "anti-vax" and other issues as part of its health integrity campaign.

Google: the Google reports vary in terms of their narration on "integrity of service" issues. In the baseline report and again in January, Google highlights its work under "Project Shield" in helping politicians, parties, journalists and others protect themselves against distributed denial of service (DDOS) attacks; the February reports and afterward focus more on efforts to promote quality newsfeeds on both Google News and YouTube, as well as takedowns of YouTube channels that don't meet Google policies for misrepresentation, spam, misleading content, and impersonation:

Google: YouTube Channel Removals in Europe, by Month, 2019

Reason/Month	February	March	April	May
Spam, misleading	628,000+	1,000,000+	900,000+	860,000+
Impersonation	5,000+	2,500+	500+	600+

Twitter: In its work to prevent "coordinated manipulation" in the run-up to the European Parliament elections, Twitter profited from its review of attempts to game the system during the fall 2018 U.S. elections. Like Facebook and Google, it works in part with government agencies to identify foreign interference, although voter suppression efforts were a major concern in the U.S. Twitter only began detailing numbers of bad accounts taken down in its March report:

Twitter: Accounts Challenged for Spam, Malicious Automation and Fake Accounts

Month/Source of Challenge	Proactively Challenged by Twitter (million)	Challenged by Twitter Users
January	19.5	489,148
February	17.0	406,162
March	16.6	504,729
April	13.8	597,295
May (1-20)	9.8	344,987
Total	76.7	2,342,321

The company in addition keeps an archive of potential foreign information operations, mainly pointing to Iran, Venezuela and Russia.

Electoral Support/Public Education

All three companies report on extensive efforts to promote digital media literacy in Europe, working with various civil society groups in the member states. They also all established electoral security centers, which trained candidates and political parties on ways to protect their sites from attack and abuse, as well as in reaching out to voters.

Facebook: Facebook, like the other platforms, places a lot of emphasis in its reports on partnering with fact-checkers; by its April report, it noted it had 21 fact-checking partnerships checking content in 14 European languages. In addition to its efforts to promote civic engagement (and get out the vote) and the engagement it (and all the companies had) in EU-supported Europe-wide media literacy campaigns, Facebook launched its own digital literacy campaign about "stamping out fake news" in all 28 EU member states, working with Full Fact and other fact-checking organizations; it also worked with over 20 civil society groups to conduct digital training to 75,000 citizens in seven EU countries. It reports on separate programs in countries such as Poland and Sweden, and notes that it will work with the German national newspaper Die Zeit to launch a major digital literacy program in June. It claims to have trained over 400 journalists in techniques to identify fake news stories.

Google: The company's reports highlight its efforts to protect European citizens from disinformation and to promote quality news content, mainly through Google News Lab. As of May, it had provided detailed training to nearly 6,000 European journalists in 27 of the 28 member states on news story verification techniques; helped launch FactCheck EU; provided security training to nearly 3,000

politicians and journalists; provided social media literacy training to over a million EU citizens; and promoted numerous voting, candidate and political party quick information pages.

Twitter: Among other things, Twitter has launched a new global partnership with UNESCO on media and digital literacy, which includes a series of resources to detect disinformation.

Research:

In addition to the Ad Transparency and Political Ad Transparency reports established by all three companies, Facebook and Twitter also report on other initiatives in Europe to promote and support research into the impact of social media on the public debate. Google did not spell out specific research-oriented work in Europe.

Facebook: Facebook, which in September 2018 established a European advisory committee for its Social Science One program to facilitate researcher access to its data, noted in its monthly reports that it:

- in April provided researchers from 60 universities from 30 academic institutions in 11 countries (including six EU member states) access to privacy-protected data under its Social Science One program;
- in May awarded grants for 19 research proposals on its content policies, including to four European universities;
- published in May the "audit" of the independent Data Transparency Accountability Group of its community standards and takedown activities.

Twitter:

- is actively engaging researchers to evaluate privacy and security changes to its "application program interface" (API, which is generally recognized as being relatively open for researchers);
- noted that its Potential Foreign Information Operations archive was reportedly accessed by over 13,000 researchers in Europe during the first five months of the year.

Notes

[1] Senior Fellow, German Marshall Fund of the United States. An earlier version of this paper was prepared for the May 2019 meeting of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression; it has been updated to reflect subsequent reporting by the signatories to the Code and revised to reflect thoughts provoked during that discussion as well as through subsequent research. The views expressed in the paper remain, however, those of the author alone.

[2] The first case of a sustained cyberattack against Estonia was in 2007, see, e.g., Damien McGuinness, How a Cyber Attack Transformed Estonia, BBC News, 27 April 2017.

[3] Statement by Commission Vice President Ansip at the European Parliament, Strasbourg, in the Plenary Debate, "Hate Speech, Populism and Fake News on Social Media – Towards an EU Response," European Commission, April 5, 2017.

[4] A little over a month following the May 7, 2017 publication in The Guardian of Carole Cadwalladr's story "The Great British Brexit Robbery: How Our Democracy was Hijacked," which reported on the role of Cambridge Analytica in the Brexit referendum.
[5] The UK government has since published an Online Harms White Paper (April 2019) which proposes an independent regulatory authority to oversee platform compliance with a "duty of care" embodied in codes of practice, including on disinformation, while the French government is also proposing a regulator to ensure platforms enforce their own terms of service/community standards.
[6] This presumably includes tweets by politicians.
[7] Hungary, Poland, Romania, Malta and Greece are all well below the EU averages in terms of trust in radio, tv, print news and online news, but above the average in terms of trust in social media and messaging aps. Eurobarometer, Fake News and Information Online, Flash Report 464, April 2018.
[8] The Opinion of the Sounding Board is available as a downloadable PDF here. Sounding Board signatories included: Grégoire Polad, Association of Commercial Television in Europe; Vincent Sneed, Association of European Radios; Oreste Pollicino, Bocconi University; Monique Goyens, Bureau Européen des Unions de Consommateurs; Ravi Vatrapu, Copenhagen Business School; Nicola Frank, European Broadcasting Union; Ricardo Gutiérrez, European Federation of Journalists; Marie de Cordier, European Magazine Media Association | European Newspaper Publishers' Association; Angela Mills Wade, European Publishers' Council; Alexios Mantzarlis, International Fact-Checking Network; Wout van Wijk, News Media Europe; Bilyana Petkova, Yale University
[9] This page on the Commission Disinformation website contains links through to all the individual monthly reports.
[10] European Commission, Joint Communication (with the EU External Action Service) to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of Regions: Report on the Implementation of the Action Plan Against Disinformation, June 14, 2019, footnote 11.
[11] Ibid. See also Alex Romero, Europe's Parliamentary Elections in the Digital Ecosystem, Disinfo Portal, July 12, 2019, updated July 15, for detailed analysis of 898 million posts across a wide range of digital media by over 95 million users in France, Germany, Italy, Poland and Spain, where one key observation is that a very small number of accounts (between 0.05 and 0.16 percent of users, mainly associated with political groups on the far left and far right, generated between 9.5 and 11 percent of activity in the five countries, mainly on socially divisive issues.
[12] Guy Rosen, An Update on How We Are Doing At Enforcing Our Community Standards, Facebook blog, May 23, 2019.
[13] For more on this problem of "over-removals," see e.g., Daphne Keller, Empirical Evidence of "Over-Removal" by Internet Companies Under Intermediary Liability Laws, Stanford University Law School Center for Internet and Society, October 12, 2015, as well "Facts and Where to Find Them: Empirical Research on Internet Platforms and Online Speech," unpublished essay, September 2018 version.
[14] See the open letter signed by 11 organizations as well as 71 researchers, Facebook and Google: This is What an Effective Ad Archive API Looks Like, The Mozilla Blog: Dispatches from the Internet frontier, March 27, 2019.
[15] Alex Krasodomski-Jones et al., Warring Songs: Information Operations in the Digital Age, Demos, Center for the Analysis of Social Media, May 2019.
[16] Jakub Kalensky, Russian Disinformation Attacks on Elections: The Case of Europe, Testimony before the Foreign Affairs Subcommittee on Europe, Eurasia, Energy and the Environment, U.S. House of Representatives, July 16, 2019, Disinfo Portal, July 17, 2019. In his testimony, Kalensky points out that the EU's East StratCom Task Force, where he previously worked, estimates that Russian disinformation activities doubled in the first half of 2019 compared to the same period the year before.
[17] See, e.g., Institute for Strategic Dialogue (ISD), Response to Online Harms White Paper Consultation, which notes, inter alia, "Harms and illegal activities are conducted through an extremely broad spectrum of technology platforms and services, as evidenced in ISD's extensive research on disinformation and extremist or terrorist use of the internet. The wide scope of platforms that would be implicated in the duty of care is, in principle, a necessity to comprehensively and sustainably address the evolving tactics of purveyors of online harm, who do not act solely on the few, largest technology platforms, but instead use an entire ecosystem of platforms to conduct harmful activity. A focus on just a few large platforms would be limited: the focus of improving content moderation approaches on a few large platforms over the past three years has led to a platform migration of many purveyors of hate speech, extremism, terrorist content and disinformation away from large platforms to smaller platforms with little or no oversight, limited or no Terms of Service (e.g. Gab), or in some cases, any appetite or intent to respond to online harms (e.g. 8chan). A limited focus on the few largest platforms would simply accelerate this phenomenon."
[18] Commission President-elect Ursula von der Leyen, A Union that Strives for More: My Agenda for Europe, European Commission, 16 July, 2019: "A new Digital Services Act will upgrade our liability and safety rules for digital platforms …" page 13.
[9] This page on the Commission Disinformation website contains links through to all the individual monthly reports.

[20] See footnote 14 supra for the online open letter published by Mozilla critiquing the transparency reports discussed here and below under "Political Advertising."

TRANSATLANTIC WORKING GROUP

Actors, Behaviors, Content: A Disinformation ABC

Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses†

Camille François
Graphika and Berkman Klein Center for Internet & Society at Harvard University[1]

September 20, 2019

Contents

Introduction

As the historic phenomenon of propaganda[2] unfolds today in a variety of social-media manifestations, a plethora of terms has emerged to describe its different forms and their implications for society: "fake news," online disinformation, online misinformation, viral deception, etc.[3] The speed and scale at which disinformation is now able to spread online has led to mounting pressure on regulators around the globe to address the phenomenon, yet its multifaceted nature makes it a difficult problem to regulate. Effective remedies must take into account the different vectors of contemporary disinformation and consider the multiplicity of stakeholders, tradeoffs in different approaches, disciplines, and regulatory bodies able to meaningfully contribute to responses.

† One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Major technology platforms have invested in better responses to disinformation, notably by adapting their community guidelines or terms of service. Observed through the lens of platform enforcement, "disinformation" breaks down into a number of different violations manifest on different products which are enforced by distinct teams. This points to a key concern with regard to the current industry responses to viral deception: while disinformation actors exploit the whole information ecosystem in campaigns that leverage different products and platforms, technology companies' responses are mostly siloed within individual platforms (if not siloed by individual products!).

This concise "ABC" framework doesn't aim to propose one definition or framework to rule them all, but rather seeks to lay out three key vectors characteristic of viral deception[4] in order to guide regulatory and industry remedies. Manipulative **actors**, deceptive **behaviors**, harmful **content**: each vector presents different characteristics, difficulties, and implications. Unfortunately, they are also often intertwined in disinformation campaigns, suggesting that effective and long-term approaches will need to address these different vectors with appropriate remedies.

This "ABC" also seeks to reconcile approaches throughout applicable disciplines (e.g., cybersecurity, consumer protection, content moderation) and stakeholders. While the public debate in the U.S. has been largely concerned with **actors** (who is a Russian troll online?), the technology industry has invested in better regulating **behavior** (which accounts engage in coordinated and inauthentic behavior?) while governments have been most preoccupied with **content** (what is acceptable to post on social media?).

"A" is for Manipulative Actors

"On the Internet, nobody knows you're a ~~dog~~ Russian military operative."[5]

The Russian disinformation campaign targeting the U.S. 2016 presidential election[6] has brought to the public's attention how keen certain government actors were to leverage social media to manipulate and influence audiences at home and abroad by engaging in information operations. It has also painfully brought to light the lack of government and industry preparedness and proactivity in the face of these threats. The cybersecurity sector, which bears the brunt of detecting these threat actors and preventing their nefarious activities, had been most focused on protecting physical networks and not enough on detecting those actors on social media networks. Facebook's April 2017 white paper on the issue of information operations (which also marks the first in-depth acknowledgement of this problem by a large technology platform) makes this point clearly and acknowledges that the Facebook cybersecurity team had to expand its scope to appropriately respond to this threat: "We have had to expand our security focus from traditional abusive behavior, such as account hacking, malware, spam and financial scams, to include more subtle and insidious forms of misuse, including attempts to manipulate civic discourse and deceive people."[7]

Manipulative actors, by definition, engage *knowingly* and with clear intent in viral deception campaigns. Their campaigns are *covert*, designed to obfuscate the identity and intent of the actor orchestrating them. Throughout the technology industry, detection and enforcement of this vector of viral deception campaigns rely on the cat-and-mouse game of a) identifying threat actors willing and able

to covertly manipulate public discourse and b) keeping those actors from leveraging social media to do so,[8] as they refine their strategies to evade detection.

Because this detection practice has its roots in the cybersecurity realm, terms of service and community guidelines do not always address these issues, or provide a clear basis to support detection and enforcement efforts against manipulative actors. Precedents in this area include platform rules laying out specific actors who are prevented from using the services (e.g.,. Foreign Terrorist Organizations[9]), but it is worth noting that no major platform to date has included language in its terms of service explicitly prohibiting governments from covertly using its services to conduct influence campaigns.[10] Setting an industry precedent, in August 2019, following investigations disclosing that Chinese State-controlled media leveraged Twitter advertising to promote content critical of pro-democracy protests in Hong Kong, Twitter announced that it would no longer allow "State-controlled media" to use its advertising products.[11] The state-controlled media entities can continue to remain "organic users" (meaning normal and/or verified accounts on the Twitter platform), but their ability to use ads to reach users who are not already following them is now restricted. In doing so, Twitter will likely face difficulty determining which entities are "taxpayer funded entities" and "independent public broadcasters" allowed to use the advertising services vs. "state-controlled media (...) financially or editorially controlled by the state" prohibited from doing so. States have also used a variety of techniques to conceal their direct involvement in seemingly independent online media properties: the Kremlin-controlled Baltnews network[12] and the Iranian-controlled IUVM[13] network are good illustrations.

Note that this problem has little to do with "banning" anonymity or pseudonymity online: both serve important purposes in protecting vulnerable voices and enabling them to participate in critical conversations.[14] Banning anonymity/pseudonymity would prevent such participation while doing little to prevent sophisticated and well-funded actors from exploiting this vector. The deceptive actors we are concerned with here are well-funded military and intelligence apparatus or campaign apparatus, not "somebody sitting on their bed that weighs 400 pounds," as President Trump famously characterized the anonymous troll. Clint Watts describes these figures as "Advanced Persistent Manipulators,"[15] a moniker that stresses the parallels and overlaps between the actors engaged in information operations[16] and hacking:[17]

> Similar to the challenge APT[18] actors have posed to information and cyber security professionals, social media companies now face malign actors that can be labeled as Advanced Persistent Manipulators (APMs) on their platforms. These APMs pursue their targets and seek their objectives persistently and will not be stopped by account shutdowns and platform timeouts.... They have sufficient resources and talent to sustain their campaigns, and the most sophisticated and troublesome ones can create or acquire the most sophisticated technology.[19]

Since 2017, we have seen multiple examples of viral deception campaigns whose *primary* vector is a deceptive actor. Notable examples include false persona "Guccifer 2.0"[20] used by the GRU, false identities tying back to the Islamic Republic of Iran Broadcasting and operating on multiple platforms,[21] and Facebook's December 2018 takedown of accounts in Bangladesh that were found to be misrepresenting their true identity and attempting to mislead voters ahead of the elections.[22]

Governments also have a role to play in detecting and mitigating harms caused by manipulative actors online, although defining the contours of government action in this space remains a largely unexplored policy question. Around the U.S. 2018 midterms elections, for instance, the U.S. government led actions to detect and share relevant information on manipulative actors with the technology sector[23] and to disrupt and deter these actors from engaging in information operations.[24]

"B" is for Deceptive Behavior

"On the Internet, nobody knows you're a ~~dog~~ bot army."

Deceptive behavior is a fundamental vector of disinformation campaigns: it encompasses the variety of techniques viral deception actors may use to enhance and exaggerate the reach, virality and impact of their campaigns. Those techniques run from automated tools (e.g., bot armies used to amplify the reach and effect of a message) to manual trickery (e.g., paid engagement, troll farms). At the end of the day, deceptive behaviors have a clear goal: to enable a small number of actors to have the *perceived impact* that a greater number of actors would have if the campaign were organic.[25]

Interestingly, while there are significant differences in the various disinformation definitions and terms of service applicable to the issue among technology companies, the focus on *deceptive behavior* appears to be a clear convergence point throughout the technology industry.

Google's definition of disinformation, as laid out in its February 2019 White Paper on "How Google Fights Disinformation," points to those deceptive behaviors as a core vector of how disinformation affects Google's platforms:

> We refer to [...] deliberate efforts to deceive and mislead using the speed, scale, and technologies of the open web as "disinformation."[26]

In Facebook's case, deceptive behavior is mostly defined through the "Coordinated Inauthentic Behavior"[27] policy, which has led to numerous takedowns since it was implemented in 2018.[28] While Facebook has shared records and data points regarding the content and accounts taken down for their participation in "coordinated and inauthentic behavior," enforcement in this realm remains opaque throughout the major technology companies.

While the detection and mitigation techniques in this area can be similar to spam detection, an area generally opaque for the public and regulators and not subject to much public scrutiny, the free speech implications of taking down *content* and social media *accounts* (especially political content during election cycles) justify much higher scrutiny of these practices. Relevant questions to technology platforms in this area include:

- Applicable rules: Which are the applicable policies set forth by the platform to address deceptive behaviors on their products?
- Enforcement: What enforcement options are available to the platforms to take action against accounts and content that violate the rules on deceptive behavior? Platforms generally acknowledge a range of options from content demotion to account suspension, although those enforcement options are rarely spelled out for users or made clear for users affected.

- Detection and prioritization: Which teams are effectively in charge of detecting deceptive behaviors, how much of this detection relies on machine learning classifiers (and which ones?), and how does prioritization of potential issues and focus areas work at the platform level?
- Transparency: How will affected users (including good faith actors mistakenly engaging in deceptive behaviors, consumers of information spread by deceptive behavior, bad faith actors seeking to best understand what telltale signs trigger enforcement, etc.) be notified when action is taken against content or accounts? Can those decisions be appealed, and if so, how? Will the platform share transparency metrics regarding its enforcement of rules relative to distortive behavior, both at the annual and the aggregate level (through the existing mechanism of Transparency Reports) and through press releases published when enforcement happens?
- Product vulnerabilities and changes: When deceptive behaviors exploit vulnerabilities in platforms and products, what changes are made to address them?[29]

The industry's lack of proactivity in tackling some of these campaigns and growing public anxiety about disinformation have led regulators to craft frameworks to specifically address deceptive behavior. California's "Bot Law," for instance, is a clear attempt to regulate deceptive behavior on social media:

> It shall be unlawful for any person to use a bot to communicate or interact with another person in California online, with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivize a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election. A person using a bot shall not be liable under this section if the person discloses that it is a bot. The disclosure required by this section shall be clear, conspicuous, and reasonably designed to inform persons with whom the bot communicates or interacts that it is a bot.[30]

The "Manipulative Actor" and "Deceptive Behavior" vectors are particularly challenging to address through effective regulatory frameworks because of the dramatic asymmetry of information between the platforms targeted by these campaigns and the rest of the world. While open-source investigation techniques and a few available tools allow others to scrutinize online activity for campaigns run by a manipulative actor or using deceptive techniques, it is undeniable that platforms have much more visibility into those issues than external researchers and stakeholders. Some platforms' community standards or terms of service either indirectly prevent the type of external research that may lead to detecting and exposing distortive behaviors (e.g., when existing and important safeguards also prevent researchers from collecting the data they'd need to analyze distortive behaviors) or directly seek to prevent it (e.g., with rules explicitly preventing the use of data in order to perform detection of deceptive behavior).

Finally, some of the platforms' own systems may actually enhance those deceptive behaviors by disinformation actors: algorithmic reinforcement is a core concern in this area.[31] While anecdotal evidence suggests machine learning based recommendations systems may easily be gamed into promoting campaigns "boosted" by adversarial distortive behavior, the difficulties discussed above

with regard to external research have prevented more systematic examinations of these issues throughout the various platforms.

"C" is for Harmful Content

"On the Internet, nobody knows you're a ~~dog~~ deepfake."

Finally, it is sometimes the case that the content of posts and messages justifies classifying a campaign as an instance of viral deception. Content is the most visible vector of the three: while it is difficult for an observer to attribute messages to a manipulative actor or to observe behavioral patterns across a campaign, every user can see and form an opinion on the content of social media posts. This is likely why regulators have focused on content aspects when regulating disinformation.

This vector calls for detection and enforcement strategies in the realm of content moderation[32]. Unfortunately, regulatory and legal frameworks often struggle to properly define categories of "harmful content" they seek to regulate (see ongoing debates about the definitions of "violent extremism," "hate speech," "terrorist content," etc.) or to properly take into account that a lot of the speech they consider to be "harmful" is protected under human rights law. Governments' appetite to regulate viral deception through the content lens risk further eroding protections to freedom of expression online.

The intersection of harmful content and disinformation campaigns can manifest in several ways:

- Entire categories of content can be deemed "harmful" because they belong to the realm of viral deception, e.g., health misinformation.[33]

Technology platforms have so far mostly proposed to address the categories of content deemed most "harmful" for their disinformation nature by adding context for users alongside the content, such as "flags" or "fact-checking" content. Some platforms though have taken a more radical route by banning entire categories of disinformation content from their services.

Photo-sharing platform Pinterest, for instance, takes action against harmful medical information shared on its platform. Its "Health Misinformation" policy reads:

> "Pinterest's misinformation policy prohibits things like promotion of false cures for terminal or chronic illnesses and anti-vaccination advice. Because of this, you're not allowed to save content that includes advice where there may be immediate and detrimental effects on a Pinner's health or on public safety."[34]

- The content of a campaign itself (not its diffusion mechanism) can be manipulated to deceive users and therefore belong to the realm of "disinformation" (e.g., use of manipulated media on the range from "deepfakes" to "cheap fakes"[35]).

- "Harmful content" can be promoted by deceptive actors or by campaigns leveraging distortive behaviors (e.g., "troll farms amplifying harassment campaigns").

It should indeed be noted that viral deception campaigns whose primary vector is a deceptive actor or distortive behavior can participate in amplifying other types of harmful content categories, such as hate speech, harassment, and violent extremism.

Conclusion and recommendations

Viral deception campaigns spread across platforms and through three core vectors: manipulative actors (A), deceptive behavior (B) and harmful content (C). As such, they represent a complex and multifaceted problem for policy makers and regulators to address. This "ABC" framework therefore offers a few modest recommendations for policy makers and regulators navigating this maze:

- Each dimension matters. Regulatory efforts focused on viral deception tend to exaggerate the role of harmful content: balanced approaches will consider how manipulative actors (both foreign and domestic) and deceptive behaviors contribute to the problem.
- Each dimension comes with its own set of challenges, tradeoffs, and policy implications. Specific disciplines may be necessary and/or best suited to address each of them. For instance, cybersecurity (and threat intelligence in particular) is a core component of how manipulative actors get detected; how the resulting signals get shared across the industry and with the relevant parties (researchers, public institutions) is a key policy question. Consumer protection frameworks (and stakeholders) may be ideally situated to help regulate deceptive behavior issues. Policies and regulatory frameworks that center around one type of remedy only (such as content takedowns*)* are insufficient.
- On a final (and related) note, Manipulative Actors (A) and Deceptive Behaviors (B) are dimensions on which the information asymmetry between the technology platforms on which this activity unfolds and the rest of the stakeholders in the debate is immense. How to ensure that the public, media, and policy stakeholders are able to meaningfully analyze both the issues and potential impacts of remedies in place is a fundamental question in this space.

Appendix: Examples of Disinformation Campaigns Spanning the Three Vectors

- A Disinformation Campaign in the Philippines (Facebook)

On March 28, 2019, Facebook removed 200 pages, groups and accounts engaged in "coordinated inauthentic behavior" on Facebook and Instagram in the Philippines. Facebook's press release[36] highlights the manipulative actor along with the deceptive behavior elements of the campaign:

> We're taking down these Pages and accounts based on their behavior, not the content they posted. In this case, the people behind this activity coordinated with one another and used fake accounts to misrepresent themselves, and that was the basis for our action.

Follow-up analysis highlights that the content taken down by Facebook in this campaign did contain "harmful content," notably in the form of hate speech and manipulated media (Photoshopped images of politicians in wheelchairs enticing viewers to question the health of candidates).[37]

- The Russian Internet Research Agency's "Columbia Chemical" Campaign (Twitter)

On September 11, 2014, a set of seemingly uncoordinated Twitter accounts engaged in disseminating news of a chemical incident and toxic fumes in the city of St. Mary Parish in Louisiana. Along with the social media campaign, videos of the "incident" were uploaded and officials and media were contacted by available channels with an alarming messaging – "Take shelter!" – and links to a dedicated website (www[.]columbiachemical[.]com).[38]

It wasn't long until officials realized that the campaign, with its false images of the incident and alarming messages, constituted harmful content – "a hoax," as it was initially described. It was later made clear that the accounts used to spread the content were coordinated to give the impression of a mounting local panic, using distortive behavior to create the illusion of a spontaneous wave of local panic.

It took a few more years for the major technology platforms and the U.S. Government to provide a final attribution on those accounts, confirming that the Internet Research Agency troll farm in Saint Petersburg was indeed the actor operating the accounts.[39]

Notes

[1] Camille François works on cyber conflict and digital rights online. She is Chief Innovation Officer of Graphika, where she leads the company's work to detect and mitigate disinformation, media manipulation and harassment in partnership with major technology platforms, human rights groups and universities around the world. She also is an affiliate of Harvard University's Berkman Klein Center for Internet & Society. An earlier version of this paper was presented as the second meeting of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression, in Santa Monica, Calif., on May 9-12, 2019. The author thanks her colleagues in the working group for their engagement with this work during the sessions. Their feedback and encouragement greatly benefited this final paper.

[2] See for instance: Tworek, Heidi JS. News from Germany: The Competition to Control World Communications, 1900–1945. Harvard University Press, 2019.

[3] For a thoughtful typology of the different aspects of the phenomenon, see for instance Claire Wardle and Hossein Derakhshan's "Information Disorder" framework: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

[4] Viral deception here is used as an umbrella term for the multiple facets of contemporary disinformation online, see Jamieson, Kathleen Hall. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press, 2018.

[5] I hope readers will forgive this 2019 edit to the famous cartoon published by Peter Steiner in the New Yorker on July 1993.

[6] See the Mueller Report: https://www.justice.gov/storage/report.pdf

[7] Jen Weedon, William Nuland and Alex Stamos, "Information Operations and Facebook", v.1: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

[8] For an example of a takedown solely motivated by the actor behind the content, see Facebook's "The IRA Has No Place On Facebook" post on April 3, 2018: "We removed this latest set of Pages and accounts solely because they were controlled by the IRA — not based on the content." https://newsroom.fb.com/news/2018/04/authenticity-matters/

[9] See for instance Microsoft's "*Approach to Terrorist Content*" statement (published May 20, 2016), which notes that "there is no universally accepted definition of terrorist content" and that Microsoft relies on organizations listed in the Consolidated United Nations Security Council Sanctions List to define and take action against terrorist content posted on its platforms: https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/

[10] There are, however, multiple policies that indirectly cover aspects of these campaigns. An example with Google Ads' "Misrepresentation Policy": https://support.google.com/adspolicy/answer/6020955?hl=en

[11] In 2017, Twitter had similarly banned Russian State-controlled media Russia Today and Sputnik from using their advertising products (https://blog.twitter.com/en_us/topics/company/2017/Announcement-RT-and-Sputnik-Advertising.html). The August 2019 policy extends this ad-hoc remediation done in the wake of the investigation

regarding the Kremlin's election interference efforts on social media to all of "state-controlled" media: https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html

[12] See the Aug. 29, 2019 BuzzFeed investigation, "This Is How Russian Propaganda Actually Works in the 21st Century": https://www.buzzfeednews.com/article/holgerroonemaa/russia-propaganda-baltics-baltnews

[13] See for instance DFRLab's "In Depth: Iranian Propaganda Network Goes Down," March 26, 2019, https://medium.com/dfrlab/takedown-details-of-the-iranian-propaganda-network-d1fad32fdf30

[14] For an examination of how manipulative actors use "pseudoanonymity" to "impersonate marginalized, underrepresented, and vulnerable groups to either malign, disrupt or exaggerate their cause," see Friedberg and Donovan's piece in the MIT JODS: https://jods.mitpress.mit.edu/pub/2gnso48a

[15] Clint Watts, "Advanced Persistent Manipulators", Feb. 12, 2019: https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/

[16] For a global inventory of actors organized for social media manipulation, see: Bradshaw, Samantha, and Philip Howard. "Troops, trolls and troublemakers: A global inventory of organized social media manipulation." (2017).

[17] See also "False Leaks: A Look at Recent Information Operations Designed To Disseminate Hacked Material," Camille Francois, CYBERWARCON 2018. Video: https://www.youtube.com/watch?v=P8iXN8j4gMk

[18] APT here refers to Advanced Persistent Threat, a term commonly used in the threat intelligence industry to describe State-sponsored and state-affiliated groups engaged in hacking operations. See: https://en.wikipedia.org/wiki/Advanced_persistent_threat

[19] Clint Watts, "Advanced Persistent Manipulators," Feb. 12, 2019: https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/

[20] Guccifer is a social media persona who claimed to be the hacker who hacked the Democratic National Committee in 2016, and who used this deceptive identity to engage WikiLeaks and the media. The account was in reality operated by Russian military intelligence: https://en.wikipedia.org/wiki/Guccifer_2.0

[21] See for instance Google's Kent Walker update on action taken against IRIB and broader State-Sponsored activity on Google's products: https://blog.google/technology/safety-security/update-state-sponsored-activity/

[22] https://newsroom.fb.com/news/2018/12/take-down-in-bangladesh/

[23] See for instance reporting by the Associated Press, "Facebook blocks 115 accounts ahead of US midterm elections", Nov. 6, 2018, https://www.apnews.com/19aabf8ba7b6466b859f4d0afd9e59be. The AP reports: "Facebook acted after being tipped off Sunday by U.S. law enforcement officials. Authorities notified the company about recently discovered online activity "they believe may be linked to foreign entities."

[24] See Ellen Nakashima's reporting in the Washington Post, "U.S. Cyber Command operation disrupted Internet access of Russian troll factory on day of 2018 midterms", Feb, 26, 2019, https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html

[25] I am borrowing here from a definition my colleagues and I have used to frame detection techniques. See Francois, Barash, Kelly: https://osf.io/aj9yz/

[26] "How Google Fights Disinformation," Feb. 2019, available at: https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf

[27] See "Coordinated Inauthentic Behavior Explained," https://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

[28] A blog post entitled "Removing Bad Actors On Facebook", from July 2018, seems to be the first public reference to "coordinated and inauthentic behavior": https://newsroom.fb.com/news/2018/07/removing-bad-actors-on-facebook/

[29] An example of a product change directly motivated by a platform's need to tackle distortive behaviors on its products can be found in the January 2019 YouTube announcement: ""To that end, we'll begin reducing recommendations of borderline content and content that could misinform users in harmful ways": https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html

[30] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

[31] See for instance former YouTube engineer Guillaume Chaslot's project regarding algorithmic reinforcement of fringe and harmful views on YouTube: https://algotransparency.org/methodology.html

[32] For an in-depth discussion of the various issues plaguing the content moderation industry, see Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. or Gillespie, Tarleton. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[33] This is a good place for a quick reminder of the differences between misinformation and disinformation. Dictionary.com, which made "misinformation" the word of the year in 2018, defines it as "false information that is spread, regardless of whether there is intent to mislead." It describes disinformation as "deliberately misleading or biased information; manipulated narrative or facts; propaganda".

[34] https://help.pinterest.com/en/article/health-misinformation

[35] See Britt Paris and Joan Donovan, "Deep Fakes and Cheap Fakes", published Sept. 18th 2019 by the Data & Society Research Institute, https://datasociety.net/output/deepfakes-and-cheap-fakes/
[36] https://newsroom.fb.com/news/2019/03/cib-from-the-philippines/
[37] https://medium.com/graphika-team/archives-facebook-finds-coordinated-and-inauthentic-behavior-in-the-philippines-suspends-a-set-d02f41f527df
[38] Adrian Chen's 2015 account in The New York Times Magazine is the first public account of this campaign: https://www.nytimes.com/2015/06/07/magazine/the-agency.html
[39] See for instance the reports commissioned by the Senate Select Intelligence Committee regarding the IRA's online activity targeting the USA: https://comprop.oii.ox.ac.uk/research/ira-political-polarization/

TRANSATLANTIC WORKING GROUP

A Cycle of Censorship:

The UK White Paper on Online Harms and the Dangers of Regulating Disinformation†

Peter Pomerantsev, Senior Fellow, Institute of Global Affairs,
London School of Economics and Political Science[1]

October 1, 2019

Introduction

This document contains two parts. The first is a summary of the UK Government Online Harms White Paper, including an overview of the arguments around it, responses to it and associated proposals by UK organisations in this field.

The second part proposes a way for the UK Government to reframe the challenge of "disinformation" as currently formulated in the White Paper. It argues that the current approach to combatting "disinformation" risks reinforcing a cycle of censorship worldwide – when the UK's publicly avowed role is to support freedom of expression globally. Another way forward is possible, one that strengthens the UK's commitments to upholding democratic values and human rights while combatting online deception.

The views and opinions expressed in this article are solely those of the author.

Contents

---

† One in a series: A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Part 1: White Paper Summary and Overview of Responses

After a year of heightened discussion in the UK about the dangers of hostile state online interference in UK democratic processes, non-transparent political campaigns by local actors, harassment of MPs and bullying of children as well as the persistent presence of violent extremism, the UK Government published an Online Harms White Paper[2] in March 2019. The White Paper attempts to find a middle way between self or no regulation on the one hand, and the imposition of liabilities for every piece of content on the other. It proposes a model of regulation that focuses on mitigating what it calls "harms" and putting systems in place to minimize "risk." So instead of being liable for each piece of content, companies are responsible for maintaining mechanisms to protect people and "society." Tech companies will be expected to follow codes of conduct and statutory "duties of care," enforced by an independent regulator.

The White Paper is a self-conscious attempt to approach internet regulation in an innovative way, and for the UK to take a leading global role on this issue. It is refreshing in its focus on the architecture of the internet, though it has also left many questions unanswered about what exactly an "online harm" is, let alone how one would mitigate it. The White Paper has had a striking number of responses in mainstream media as well as among the expert community, with a lively exchange of ideas between a variety of critics and defenders. It has shown structural differences between freedom of expression and human rights advocates on the one hand, and politicians' desires to reform the online space. The lack of consensus is worrying as it betrays fundamental lack of clarity about the UK's vision for the future of the information space, clearly a critical area for democracy.

The analysis below first addresses the main ideas in the White Paper and the overall criticisms of them. While the White Paper is the main focus, I also look at studies and proposals that have influenced the debate in the UK:

- The Digital, Culture, Media and Sport Committee's final report on Disinformation and 'fake news'[3]
- The UK broadcasting regulator, OFCOM, report on Addressing Harmful Online Content[4]
- Carnegie UK's proposal for Internet Harm Reduction[5]
- Full Fact (UK's leading fact checker) paper on Tackling Misinformation in an Open Society[6]

Summary of White Paper Proposals

Very broad scope: The Online Harms White Paper foresees creating mechanisms to regulate companies "that allow users to share or discover user-generated content or interact with each other online," e.g., "two main types of online activity": hosting, sharing and discovery of user-generated content, and facilitation of public and private online interaction between service users. No exception is allowed based on size or nature of the services. However, the framework should ensure a differentiated approach for private communication, "meaning any requirements to scan or monitor content for tightly defined categories of illegal content will not apply to private channels." The regulator will provide support to start-ups and small- to medium-sized enterprises (SMEs) "to help them fulfill their legal obligations in a proportionate and effective manner."

A new statutory duty of care will be introduced to "make companies take reasonable steps to keep their users safe and tackle illegal and harmful activity on their services." These will be fleshed out with specific codes of practice targeted to each specific type of harm (see below). However, companies will still need "to be compliant with the overarching duty of care even where a specific code does not exist," continuing to assess and respond "to the risk associated with emerging harms or technology."

Online harms: the codes of practice will cover harms "**with clear definition,"** such as child sexual exploitation, terrorist content and activity, organised immigration crime, modern slavery, extreme pornography, revenge pornography, harassment and cyberstalking, hate crime, encouraging or assisting suicide, incitement of violence, sale of illegal goods/services (such as drugs and weapons on the open internet), content illegally uploaded from prisons, or sexting of indecent images by under 18-year-olds (creating, possessing, copying or distributing indecent or sexual images of children and young people under the age of 18). They will also cover harms "**with a less clear definition,"** such as cyberbullying and trolling, extremist content and activity, coercive behaviour, intimidation, disinformation, violent content, advocacy of self-harm, and promotion of female genital mutilation. **Underage exposure to legal content** will also be in scope, such as children accessing pornography, children accessing inappropriate material (including under 13 years of age using social media and under 18 years of age using dating apps; and excessive screen time).

Online harms excluded: all harms to organisations (e.g., competition law, most cases of IP violation) and all harms suffered by individuals resulting directly from a breach of the data protection legislation and from a breach of cyber security or hacking are outside the scope of paper, as these are covered by other applicable laws.[7]

Codes of practice will set out how this duty of care should be fulfilled (through "systems, procedures, technologies and investment, including in staffing, training, and support of human moderators"). Companies that do not wish to comply with these codes will have to explain how what they will be doing will have the same or a greater impact. These codes of practice will be drafted:

(1) under the direction of an independent (existing or to be created) regulator;

(2) under the direction of the British government for codes of practice on terrorist activity or child sexual exploitation online; and

(3) in cooperation with **law enforcement for "illegal harms"** (such as incitement of violence and sale of illegal goods and services).

Compliance with this duty of care will be overseen and enforced by the regulator. This regulator will "set out expectations for companies to do what is reasonably practicable to counter harmful activity or content, depending on the nature of the harm, the risk of the harm occurring on their services and the resources and tech available to them." When assessing compliance, the approach taken will be "risk-based," "prioritising regulatory action to tackle harms that have the greatest impact on individuals or wider society" and will consider "whether the harm was foreseeable, and therefore what is reasonable to expect a company to have done."

Upload filters: When overseeing the implementation of the codes of practice, the regulator "will not compel companies to undertake general monitoring of all communications on their online services." However, the

British government considers that "there is a strong case for mandating specific monitoring that targets where there is a threat to national security or the physical safety of children."

Enforcement power: The regulator will be able to issue substantial fines, to ask for the publication of annual transparency reports, to require additional information from companies (including the use of algorithms in selecting content for users), to oversee the implementation of user redress mechanisms, to promote the development and adoption of safety technologies to tackle online harms and to "encourage access of independent researchers" to the data of companies. Additional powers to **disrupt the business activities** of a non-compliant company may be granted following the consultation, e.g., measures forcing third-party companies to withdraw any service they provide that directly or indirectly facilitates access to the services of the non-compliant company, ISP blocking and/or imposing **liability on individual members of senior management**. The regulator will be funded by industry in the medium term, and will have a "legal duty to pay due regard to innovation, to protect users' rights online."

White Paper and 'Viral Deception'

The White Paper claims the regulator will not "police truth on the internet" but mitigate the harms caused by "disinformation" and "online manipulation," which constitute "Threats to Our Way of Life":

> Our society is built on confidence in public institutions, trust in electoral processes, a robust, lively and plural media, and hard-won democratic freedoms that allow different voices, views and opinions to freely and peacefully contribute to public discourse.
>
> Inaccurate information, regardless of intent, can be harmful – for example the spread of inaccurate anti-vaccination messaging online poses a risk to public health. The government is particularly worried about disinformation (information which is created or disseminated with the deliberate intent to mislead; this could be to cause harm, or for personal, political or financial gain).
>
> Disinformation threatens these values and principles, and can threaten public safety, undermine national security, fracture community cohesion and reduce trust.

This is very vague language, but as the White Paper elaborates on "disinformation" and "online manipulation" it makes a laudable move away from obsessing over content to focus on behavior, actors and online architecture, though in a somewhat unstructured way.

Under "threats" from online disinformation, the White Paper quotes studies from the Oxford Computational Propaganda Project about the extent of organised social media manipulation, highlights the risks posed by micro-targeting and identifies the disinformation campaigns of the Russian state as particularly noteworthy. The White Paper describes the "impact" of this threat as being that people are largely not aware that algorithms define what they see online, and that their personal data, browsing history and networks play a part in this. Only 3 in 10 adults, the White Paper states, are aware of how companies collect people's data online.

"Online manipulation" seems a slightly wider category of harm than "disinformation." The White Paper argues that while "tolerance of conflicting views and ideas are core facets of our democracy" they are "inherently vulnerable to the efforts of a few to manipulate and confuse the information environment for nefarious purposes, including undermining trust. A combination of personal data collection, AI based

algorithms and false or misleading information could be used to manipulate the public with unprecedented effectiveness."

The White Paper makes a distinction between "legitimate influence" and "illegitimate manipulation." Examples of the latter in broadcasting regulation include subliminal messaging and other techniques which influence people without them being aware. Again, a recurring motif seems to be that people's lack of understanding about how and why the information environment around them is being shaped in certain ways is in itself a harm.

When it comes to fulfilling the Duty of Care on disinformation, the White Paper suggests that "companies will need to take proportionate and proactive measures to help users understand the nature and reliability of the information they are receiving, to minimise the spread of misleading and harmful disinformation and to increase the accessibility of trustworthy and varied news content."

The areas the regulator will include in a code of practice can be broken down into several categories.

i) Transparency: measures to ensure people understand what they are seeing online and why. People should know "when they are dealing with automated accounts," and there needs to be more clarity around political advertising.

This begs the question of what exactly is "political advertising." Electoral law defines electoral material as "material which can reasonably be regarded as intended to promote or procure electoral success at any relevant election." As Full Fact argue, this is too narrow a definition, and there is a need for transparency around political messaging outside of campaigns. Full Fact believe that advertising transparency requires full information on content, targeting reach and spend, which must all be provided in real time. They call for all factual claims used in political adverts to be pre-cleared, and compulsory watermarks to show the origin of online adverts.

ii) Cooperation with Fact Checkers and Boosting Authoritative News: The White Paper says companies will need to make "content which has been disputed by reputable fact-checking services less visible to users" and promote "authoritative news sources." Companies will need to promote "diverse news content, countering the 'echo chamber' in which people are only exposed to information which reinforces their existing views." It quotes the Cairncross Review, which "proposed that a 'news quality obligation' be imposed upon social media companies, which would require these companies to improve how their users understand the origin of a news article and the trustworthiness of its source."

These exhortations are, however, difficult to enact in practice. Fact-checking bodies have had a mixed experience working with technology companies. There is a problem due to the difference in scale between tech companies and fact-checkers, with fact-checkers feeling that they have no impact on how content is eventually shown online and fear they are being used as PR cover; there are also differences of opinion about which sorts of dis-, mis- and mal-information to flag.

Defining "quality news" is also difficult. Two approaches predominate: Voluntary associations that a media organisation can join, membership of which guarantees certain standards. This, however, risks leaving out the credible sites (blogs, Facebook publications in authoritarian countries, etc.) that have not joined such associations. Another approach is AI driven, and tries to automatically check for factual errors on domains or

suspicious metadata that suggests sites have been put together too quickly to be professional. Such AI-driven approaches are still highly unreliable.

iii) Algorithmic Accountability: Point 7.30 in the White Paper demolishes the current business model of at least one well-known tech company: "Companies will be required to ensure that algorithms selecting content do not skew towards extreme and unreliable material in the pursuit of sustained user engagement."[8]

The line has no further elaboration in the White Paper, but it could be of vast significance. The White Paper provides no clarity how any algorithmic oversight would work in practice. Will it entail being given access to the process of how tech companies create and adjust their algorithms? In which case how will competitive advantages and innovations be preserved? Or will there be some method of judging whether technology companies' promises to adjust algorithms are being followed?

Responses to the White Paper

The White Paper has drawn sharp criticism, which roughly falls into two camps.

The first line of criticism stresses the "chilling effect" on free speech, fearing that companies will be over-zealous in their takedowns given how tough the potential sanctions are. Index on Censorship, for example

> is particularly concerned about the duty of care. The concept is closely linked to the 'precautionary principle,' which has been widely applied in the environmental field, where it means not waiting for full scientific certainty before taking action to prevent harm. This makes sense. However, applying the precautionary principle to freedom of expression runs a high risk of legitimising censorship, especially when combined with large fines. It creates a strong incentive for online platforms to restrict and remove content.[9]

Article 19 "strongly opposes any 'duty of care' being imposed on Internet platforms. We believe a duty of care would inevitably require them to proactively monitor their networks and take a restrictive approach to content removal."[10]

In her Twitter feed Professor Lilian Edwards points out the threats to block information service providers to a country could already be in breach of ECHR rules (ECHR Article 10) and the e-Commerce Directive stating that information service providers will not be made subject to prior authorisation "or other requirements having equivalent effect."

Another line of criticism focuses on the nebulous definition of "harm" in the White Paper. Confusingly the White Paper says harms will be "evidence based," but provides no evidence of harm. The well-known right-wing columnist Toby Young[11] complains "the word 'harm' isn't defined, even though it appears in the title. That's alarming because the white paper says the new regulator will ban online material 'that may directly or indirectly cause harm' even if the content in question is 'not necessarily illegal'. As an example of what it has in mind, the government singles out 'offensive material', as if giving offence is itself a type of harm. Merely showing that it hasn't caused the complainant any tangible harm won't be sufficient, since all the regulator will need to show is that it *may* cause them *indirect* harm." If we already have legislation on illegal harms (such as child sexual exploitation), this argument runs, why do we need more?

Many critiques stress the difficulties of transferring offline to online harms. Index on Censorship argue that "although social media has often been compared to the public square, the duty of care model is not an exact fit because this would introduce regulation – and restriction – of speech between individuals based on criteria that is far broader than current law." Graham Smith, a lawyer specializing in internet law and author of the Cyberleagle[12] blog, makes the point that while it is tempting to draw a simple comparison between offline and online "duties of care," it is worth bearing in mind that the former:

- are restricted to objectively ascertainable injury;
- rarely impose liability for what visitors do to each other; and
- do not impose liability for what visitors say to each other.

Smith argues that online "harms" and "duties of care" could overstep these boundaries. Smith goes on to say that the proposed online duty of care is no duty of care at all. A "proper" duty of care, he argues, is a "legal duty owed to identifiable persons. They can claim damages if they suffer injury caused by a breach of the duty. … Occasionally a statute creates something that it calls a duty of care, but which in reality describes a duty owed to no-one in particular, breach of which is (for instance) a criminal offence." Smith quotes an environmental law in respect of waste disposal, but which is nevertheless precise "about the conduct that is in scope for the duty." He concludes that this is a mechanism "unsuited to what people say and do online."

The White Paper leaves it to the regulator to decide what exactly "legal but harmful" behavior and activity entails in practice, and what precisely would be demanded of tech companies to show they have provided "duty of care": "its very *raison d'etre* is flexibility, discretionary power and nimbleness," writes Smith. "Those are a vice, not a virtue, where the rule of law is concerned."

For supporters of regulation, one response to such criticism has been to reference British broadcasting regulation. The DCMS Parliamentary Committee on "Fake News," for example, argues that the regime for regulating broadcast content standards could be used "as a basis for setting standards for online content." The Carnegie UK paper on Internet Harm Reduction[13] quotes a House of Lords debate on a social media duty of care where Baroness Grender argued that competent regulators have had little difficulty in working out what harm means:

> "If in 2003 there was general acceptance relating to content of programmes for television and radio, protecting the public from offensive and harmful material, why have those definitions changed, or what makes them undeliverable now? Why did we understand what we meant by "harm" in 2003 but appear to ask what it is today?"

OFCOM's task in the Communications Act 2003 to which Baroness Grender refers is somewhat harder than merely harm:

> "generally accepted standards are applied to the content of television and radio services so as to provide adequate protection for members of the public from the inclusion in such services of offensive and harmful material."

OFCOM's own paper on the subject, however, is more circumspect. It highlights the difference between broadcasting and online content. Not only are volumes of content much greater online, but the public also

could have very different expectations for online content. The multinational nature of platform operators is an added complication.

The OFCOM paper argues that the priorities for online standards setting should be in protecting minors, protection from illegal content, bullying and "trolling." Setting standards in "news," however, is complex.

The fact that the government regulator is a priori at odds with the government's position on regulating political "disinformation" and "fake news" is indicative of the systemic disagreements on how to tackle this area.

Conclusion

Though an innovative approach to regulating the online space, and one that does well to go beyond ill-advised attempts to regulate all user-generated content, the White Paper has several structural faults that will need to be overcome in future drafts.

Among those are two that are especially critical for the government to address. It will need to:

- articulate more clearly how it plans to defend freedom of expression online; and
- draw clear distinctions between the regulatory framework for illegal content on the one hand, and on what it terms "harmful but legal" content on the other. Two such vastly differing categories cannot reasonably be placed under one framework.

Due to the amount of criticism directed at the White Paper, a revised version is expected at the start of 2020.

Part 2: Breaking the Cycle of Censorship

The dangers of regulating disinformation, and how the UK Government can reframe the debate to undermine dictators and strengthen democracy

In July 2019, the UK Government held the Global Conference on Media Freedom. Hosted by the British Foreign Minister, the government claimed the conference to be the "first of its kind,"[14] "part of an international campaign to shine a global spotlight on media freedom and increase the cost to those that are attempting to restrict it." Human Rights lawyer Amal Clooney gave the opening speech in front of an audience of journalists and activists from across the world, many of whom have faced vicious attacks and censorship from governments in their home countries. Clooney spoke movingly about her work as a human rights lawyer defending journalists in oppressive regimes and about how freedom of expression is in danger.[15] The conference was a high-profile statement of intent from the UK, as it plans to make media freedom one of its headline foreign policy priorities.

In the same month that the British government hosted this highly publicized conference, freedom of expression organisations were submitting their response to the government's "White Paper on Online Harms." The White Paper proposes to impose a mandatory "duty of care" on tech companies that will force them to show they are mitigating both clearly illegal activity as well as what the government terms "legal but harmful" content on their platforms, including disinformation. In their responses to the White Paper, freedom of expression groups cast the UK Government in a quite different light from the one in which it presented itself at the Global Conference on Media Freedom:

> "We have significant concerns over the scope of harms included as well as the model being proposed, and the risks that they would pose to the rights to freedom of expression and privacy," wrote Global Partners Digital in a statement that was echoed by many other groups. "We believe that the proposals, if taken forward in their current state, would likely put the UK in breach of its obligations under both international human rights law and the European Convention on Human Rights (ECHR), as incorporated into domestic law through the Human Rights Act 1998 (HRA 1998)."

How has the UK government's split policy personality on freedom of expression come about? Can anything be done to resolve it?

Part 2 of this paper explores how the manner in which the UK Government has framed the challenge of online disinformation risks damaging the very ideals it espouses in its domestic and foreign policy, and then sets out ways the UK Government can reframe its approach in order to play the global role it has set for itself as defender of freedom of expression.

The UK White Paper on Online Harms shows that the debate on regulating online "disinformation" has reached a critical fork in the road. Along one path lies an opportunity to strengthen freedom of speech, human rights and deliberative democracy for the 21st century; down another lies the risk of fundamentally misunderstanding the nature of the internet, damaging freedom of speech, and imposing a political logic and language that favors authoritarian regimes. While the White Paper has some encouraging ideas that hint at the possibility of taking the former path, its overall language and framing show that the UK Government risks stumbling, perhaps unthinkingly, down the second.

The Censorial Cycle

The UK White Paper is right to demand that tech companies take more responsibility for illegal material and behavior on their platforms. From Facebook-fueled ethnic cleansing in Myanmar to the continued circulation of child pornography online, tech companies need to comply with national and international law and with international human rights legislation especially. The White Paper is also wise in not making companies liable for every piece of content on their platforms, which is almost impossible to enforce technically, but to demand they put in place the systems to mitigate the dissemination of illegal content.

However the White Paper is on much thinner intellectual ice when it comes to the more delicate question of what it calls "harmful but legal" content, which it intends to lump together under the same "duty of care" as outright illegal material, and which includes everything from online bullying to the subject of this paper, disinformation.

"Disinformation" is not a legal concept, and clamping down on it unavoidably runs counter to international legislation on freedom of expression. As the media law scholar Damian Tambini argues,

> The central legal and constitutional problem here is that establishing new standards in a code of conduct, and introducing sanctions and fines for "harms with a less clear definition" and that are also legal, does not pass the European Convention on Human Rights free speech test according to which restrictions have to be prescribed by law, and necessary, for a legitimate aim. The requirement that restrictions should be "prescribed by law" is a safeguard against a slippery slope to censorship. Constraints on speech should not be imposed on the basis of opaque agreements between platforms and politicians – a scenario arguably left open by the White Paper – they should be subject to the constraint of parliamentary debate.[16]

As many free speech groups have noted, the extensive list of punishments companies risk being subjected to if they do not comply with the duty of care (which include fines, blocking their business from operating, even imprisonment) risks a "chilling effect" on free speech, where tech companies would rather take material down than face the risk of such punishment. Though the White Paper does make some noises about protecting freedom of expression, in practice its proposals show scant respect for it. For example, the White Paper notes the need for a "super complaints" procedure for users to demand action from technology companies when they do not abide by their Duty of Care. One might expect such a "Super Complaints Procedure" to primarily protect people whose content has been taken down to seek redress. Instead the White Paper stresses that the super complaints procedure should be used for the opposite, as a way for individuals to demand companies take material down.

But there is more at stake here than just complaints procedures around content moderation. As currently formulated the White Paper's logic and language are skewed toward suppression rather than defense of freedom of expression. The White Paper invokes an idea of speech as somehow inherently dangerous, something that people need to be protected from. This is a logic and language that will suit those authoritarian regimes, such as Russia, that aim to frame the debate around internet regulation in such a way as to legitimize censorship, to create what Russian and Chinese advocates of censorship call a "sovereign internet" where they can control content and slow down the free flow of information across borders.[17] As David Kaye, the UN Rapporteur on Freedom of Speech and a professor at the University of California, Irvine, says: "A 'rhetoric

of danger' is exactly the kind of rhetoric adopted in authoritarian environments to restrict legitimate debate, and we in the democratic world risk giving cover to that."

There is a distinct irony at work here: one of the motivations for introducing regulation in the UK has been the covert online campaigns waged by the Kremlin to influence democratic processes in the U.S. and Europe. Now the UK's response risks imposing exactly the sort of ideas for governing the internet the Kremlin is promoting.

One could argue that authoritarian regimes will enact censorship irrespective of how democracies regulate the internet. This may be the case, but the mission of democracies should be to propose an alternative regulatory vision, one that empowers and inspires in line with human rights and strengthens democratic ideals worldwide. This is especially true of the UK, which sees itself as a global leader in setting standards for media freedom and information policy. When creating regulatory proposals policy makers should therefore always be asking themselves what is the difference between an authoritarian internet and a democratic one, how do regulatory proposals strengthen the latter or at least avoid damaging it. By taking on the language and logic of authoritarian regimes, the UK Government risks reinforcing a censorial cycle: the more covert online influence campaigns authoritarian regimes such as Russia launch in democracies, the more democracies adopt a frame and policy logic these regimes favour.

We already see this cycle being perpetuated in the raft of policy proposals across the world to combat "fake news." A German law to take down "illegal" content – including blasphemy – has been quoted by Russia and Singapore as they put forward their own punitive legislation. The Singapore law is committed to fighting "fake news" and "disinformation" – but as defined by the government.[18] Journalists fear it will be used as an excuse to attack them.

In *Don't Think of an Elephant!,* the cognitive linguist George Lakoff defines winning and losing in politics as being about framing issues in a way conducive to your aims. Defining the argument means winning it. If you tell someone not to think of an elephant, they will end up thinking of an elephant. "When we negate a frame, we evoke the frame…when you are arguing against the other side, do not use their language. Their language picks out a frame – and it won't be the frame you want."[19]

But even as one rejects the "rhetoric of danger" and the negative framing the White Paper shares with authoritarian regimes, one needs to also recognize how online disinformation is qualitatively different than older forms. While producing erroneous content is not new, technology now makes it possible to disseminate content at unheard-of rates, targeted at specific audiences. As the law professor Tim Wu argues:

> The most important change in the expressive environment can be boiled down to one idea: it is no longer speech itself that is scarce, but the attention of listeners. Emerging threats to public discourse take advantage of this change… emerging techniques of speech control depend on (1) a range of new punishments, like unleashing "troll armies" to abuse the press and other critics, and (2) "flooding" tactics (sometimes called "reverse censorship") that distort or drown out disfavored speech through the creation and dissemination of fake news, the payment of fake commentators, and the deployment of propaganda robots."[20]

Speech itself, argues Wu, is being used as a "censorial weapon."

So how can one simultaneously respond to the specific challenge of digital era disinformation, while strengthening democratic ideals and freedom of expression?

Frame 1: Don't Mention It

To achieve the minimal aim of not damaging freedom of expression with its regulatory proposals, the UK Government can simply avoid regulating legal (if untrue) speech as a category in and of itself, and instead expand on existing legislation sector by sector and environment by environment to deal with disinformation in specific contexts. This is a sort of framing by omission, where the "disinformation" question is not dealt with as a separate category, but becomes a subset of other legislation.

An obvious place to start is electoral advertising. Current regulations around transparency and accuracy of political advertising and election integrity focus on traditional print and broadcast media. These must be updated to address the new reality of online political microtargeting, where adverts are created in the millions with different messages aimed at niche audiences. There needs to be a legal requirement to create an easily searchable repository for all election-related ads in real time that clearly shows who has paid for them, to whom they are targeted, and which of a person's data is used to target them. Moreover, as the Coalition for Reform of Political Advertising and Incorporated Society of British Advertisers propose, all factual claims in the political ads should be pre-cleared and the ads should be watermarked to show their origins.[21] As political parties in the UK pulled out of regulation from the Advertising Standards Authority, essentially, in the neat phrase of Full Fact, "political parties have chosen to hold themselves to lower standards than washing powder sellers," this regulatory function will have to be placed under a body such as auspices of, for example, the Electoral Commission.

The challenge, of course, is whether the current definition of "electoral ads" is sufficient. Electoral law currently defines electoral material as "material which can reasonably be regarded as intended to promote or procure electoral success at any relevant election." Just a glance at the current environment in the UK, where non-transparent online ads are being used to influence the Brexit debate as we speak, shows how online ads are being used all the time to shape political outcomes. All ads by political parties and government agencies should show full information on content, targeting reach and spend, which must all be provided in real time. This still, however, would not cover the full spectrum of political ads, such as issue-based ads by civic groups, proxies and allies. Indeed, given there is no settled scope of what a "political" ad is, ultimately all paid-for content should have this level of transparency attached.

Elsewhere existing legislation on public health could be drawn on to ensure that people are informed of health risks propagated by inaccurate online disinformation about, for example, vaccines. As to foreign interference campaigns, where not covered by regulation around election integrity and political advertising, the most egregious cases could be addressed under national security policy. Importantly, national security policy not only comprises legislative and regulatory measures, but may also engage diplomatic channels for the resolution of foreign interference. The pushback against covert foreign campaigns might not be in the information space at all, but in asymmetric responses such as economic sanctions against hostile states.

Election, privacy, national security and public health are not an exhaustive list of sectors that need to be updated for the digital age, but they are obvious examples of how a sectoral approach would deal with specific aspects of "disinformation" without having to impose blanket bans on types of speech. As discussed, content

moderation of "disinformation" will invariably bring a collision with freedom of speech. It is also largely impractical, encouraging a "whack-a-mole" approach. Most importantly it fails to understand that information operation campaigns, such as the infamous Russian Internet Research Agency campaign in the U.S., can use neutral or even accurate content in their activity. Much of the Russian covert U.S. campaign, for example, simply supported various causes and politicians, without giving any specific accurate or inaccurate facts. Rather than deceptive content, these campaigns are marked by what Facebook calls "coordinated inauthentic behavior" or what one could term "viral deception": where the actors disguise both their true identity in order to deceive people, and where material is promoted in an inauthentic way to make it look more popular than it is.

Frame 2: From Content to Behavior

If we reframe "disinformation" as pertaining less to content and more to behaviour – here, referring to (artificial) technical means to boost the dissemination of certain content – we get away from the problem of regulating speech and onto the systemic use of technology to deceive people, from regulating statements to focusing on the use of bots, cyborgs and trolls that purposefully disguise their identity to confuse audiences; cyber-militias whose activity seems organic but who are actually part of planned campaigns full of deceptive accounts; the plethora of "news" websites free of journalistic standards that look independent but are covertly run from one source, all pushing the same agenda. The issue here is not anonymity, which is sometimes necessary to guarantee safety, but the right of people to understand how the information environment around them is being shaped. Shouldn't one have the right to know if what looks organic is actually orchestrated? How the reality one is interacting with is engineered? Shouldn't bots, to give a small example, always be clearly marked as bots?

A first legislative step in this direction has been taken in California, where bots are now forced to reveal their "artificial identity" when they are used for commercial or electoral purposes. "It's literally taking these high-end technological concepts and bringing them home to basic common-law principles," Robert Hertzberg, a California state senator who is the author of the bot-disclosure law, told the New Yorker. "You can't defraud people. You can't lie. You can't cheat them economically. You can't cheat 'em in elections."[22]

Could such transparency around online behavior be taken further? Can we imagine an online life where any person would be able to understand how the information meteorology around them is being shaped; why computer programs show you one piece of content and not another; why any ad, article, message or image is being targeted specifically at you; which of your own data has been used to try and influence you and why; whether a piece of content is genuinely popular or just amplified. Ideally such information should be instantly available in real time, so that, for example, one could click on or hover over a piece of online content and be able to immediately access its provenance. Ultimately different technological solutions will need to be found for different platforms. What matters is framing the issue in a way that demands more information, not censorship, which empowers the user. Maybe then we would become less like creatures acted upon by mysterious powers we cannot see, made to fear and tremble for reasons we cannot fathom, and instead would be able to engage with the information forces around us as equals.

Such a reframing of the "disinformation" debate to focus on uncovering deceptive behaviour and empowering a person online to understand the information environment around them takes us away from the logic and

language that authoritarian regimes promote. We are back in a framing that increases people's rights and freedoms, rather than constricting them. As David Kaye, the UN rapporteur on Freedom of Expression, told me:

> Another way to conceptualize the impact and purpose of viral deception – assuming we can define it sufficiently narrowly – is as a tool to interfere with the individual's right to information. Coordinated amplification has a strategic aim: make it harder for individuals to assess the veracity of information. It harms public debate, but it also interferes with the individual's (per A19 of the ICCPR/UDHR) "right to seek, receive & impart information and ideas of all kinds." Conceived this way, it seems to me that solutions could be aimed at enhancing individual access to information rather than merely protecting against public harm.[23]

Internet Transparency as Unifying Principle for Democracies

Focusing on disinformation as pertaining to "inauthentic" behaviour in disseminating content helps open a broader discussion about the transparency of the internet, an approach that freedom-of-expression advocacy organisations are more comfortable with than regulating content itself. As Article 19 write in their response to the White Paper:

> …the regulator should not be involved in the determination of the legality of content, but instead focus on transparency obligations and reviewing internal company processes on content moderation…
>
> Government should focus on greater transparency and accountability mechanisms in the application of companies' terms of service/community standards … digital companies should explain to the public how their algorithms are used to present, rank, promote or demote content. Content that is promoted should be clearly marked as such, whether the content is promoted by the company or by a third-party for remuneration … they should publish information about the methods and internal processes for the elaboration of community rules.[24]

This focus on "transparency" has at least three advantages.

First, it puts the focus on empowering people online, augmenting their right to receive information, rather than constricting freedom of expression.

Second, greater transparency will help balance the information field to give public interest news organisations, such as fact-checking agencies, a fighting chance.

Though there have been some tenuous efforts by social media companies to work with fact-checkers, the relationship is deeply unequal and the lack of transparency makes it hard for fact-checkers to operate with efficiency. In interviews I have conducted with fact-checkers, for instance, many complain they have little knowledge about which pieces of disinformation they should focus on, as they cannot see which pieces of disinformation are being aggressively amplified by political actors. Instead they can spend time and effort on debunking ineffectual lies. If there were enough transparency to show which pieces of disinformation are being pushed through targeted, coordinated, inauthentic campaigns, then this would signal to the fact-checkers which content to focus on.

Third, and perhaps most important in the context of this text, internet transparency can become the consensus position that democracies can agree on as a unifying principle for both regulating the internet and promoting democratic values, one that stands in robust contrast to the mix of censorship and non-transparency that define authoritarian approaches. This is a way for the UK to be both a global leader in promoting media freedom and promoting a way to regulate the internet that protects freedom of expression. When the next Global Conference on Media Freedom rolls around, the UK will be able to trumpet how its vision for the internet is in harmony with its noble support to protect journalism.

Notes

[1] Peter Pomerantsev is a Senior Fellow at the Institute of Global Affairs at the London School of Economics and Political Science, where he is Director of the Arena Initiative. An author and TV producer, he specializes in propaganda and media development, and has testified on the challenges of information war to the U.S. House Foreign Affairs Committee, U.S. Senate Foreign Relations Committee and the UK Parliament Defense Select Committee. He is the author of *This Is Not Propaganda: Adventures in the War Against Reality*, published in August 2019 by PublicAffairs, and *Nothing Is True and Everything Is Possible* (2016), which won the Royal Society of Literature Ondaatje Prize.

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

[3] https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm

[4] Ofcom, September 2018, "Addressing Harmful Online Content"

[5] Woods, Perrin. Carnegie UK, January 2019, "Internet Harm Reduction"

[6] Full Fact, 2018, "Tackling Misinformation in an Open Society"

[7] To note: UK Music has already called upon the British Government to expand the scope of the paper to protect the culture and creative industries.

[8] https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm

[9] https://www.indexoncensorship.org/2019/04/uk-government-online-harms-white-paper-shows-disregard-freedom-expression/

[10] https://www.article19.org/resources/uk-article-19-response-to-leaked-reports-on-online-harms-white-paper/

[11] https://www.spectator.co.uk/2019/04/iplod-sajid-javids-new-internet-rules-will-have-a-chilling-effect-on-free-speech/

[12] https://www.cyberleagle.com/

[13] https://www.carnegieuktrust.org.uk/publications/internet-harm-reduction/

[14] https://www.gov.uk/government/topical-events/global-conference-for-media-freedom-london-2019/about

[15] https://www.gov.uk/government/speeches/addressing-threats-to-media-freedom-amal-clooneys-speech

[16] Reducing Online Harms through a Differentiated Duty of Care: A Response to the Online Harms White Paper | FLJS https://www.fljs.org/content/reducing-online-harms-through-differentiated-duty-care-response-online-harms-white-paper

[17] https://www.hrw.org/news/2019/04/24/joint-statement-russias-sovereign-internet-bill

[18] https://www.nybooks.com/daily/2019/07/19/singapore-laboratory-of-digital-censorship/

[19] Lakoff, George. *Don't Think of an Elephant!* Chelsea Green, 2014

[20] https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete

[21] https://coinform.eu/wp-content/uploads/2019/02/Full_fact_tackling_misinformation_in_an_open_society.pdf

[22] https://www.newyorker.com/tech/annals-of-technology/will-californias-new-bot-law-strengthen-democracy

[23] https://www.the-american-interest.com/2019/06/10/how-not-to-regulate-the-internet/

[24] https://www.article19.org/wp-content/uploads/2019/07/White-Paper-Online-Harms-A19-response-1-July-19-FINAL.pdf

Design Principles for Intermediary Liability Laws[†]

Joris van Hoboken, Vrije Universiteit Brussel and University of Amsterdam[1]
Daphne Keller, Stanford Center for Internet and Society[2]

October 8, 2019

Contents

I. Introduction

The goal of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (TWG) is "to identify and encourage adoption of scalable solutions to reduce hate speech, violent extremism and viral deception online, while protecting freedom of expression and a vibrant, global internet." This goal raises two central questions:

- what are optimal *policies* with respect to hate speech, violent extremism and viral deception in the private sector?
- what is the optimal *legal framework* to promote such policies in the private sector, while protecting freedom of expression online?

As explored in this discussion paper,[3] intermediary liability (IL) frameworks provide answers that are at the intersection of these two questions. They define platforms' legal responsibilities in moderating and managing content posted by internet users. Specific intermediary liability laws, such as the U.S. Communications Decency Act of 1996, Section 230 (CDA 230), and Articles 12-15 of the EU's e-

[†] One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Commerce Directive (ECD), were put in place in response to the rise of the internet in the 1990s. They provide internet services acting as intermediaries with so-called safe harbors from liability for the activities of third parties using their services. Such laws may, for example, define or restrict the liability that can be imposed on a social media service for defamatory comments of users or on broadband providers for giving access to a website that facilitates illegal file-sharing.

There were a number of reasons why intermediary liability laws were adopted at the time. During the 1990s, intermediary service providers became the targets of litigation for the behavior of users. The open nature of these services, whose providers typically would not exercise prior control over the contents of information and communication of users, raised complex questions about the allocation of legal responsibility for harmful and/or illegal behavior.[4] Intermediaries became litigation targets and such litigation about the precise responsibilities of different intermediaries under existing laws led to legal uncertainty. These disputes raised business risks for the nascent internet service industry, caused legal fragmentation among different countries or regions, and raised concerns about the proper balance between effective remedies for harm and the protection of freedom of expression.

In response, statutory intermediary liability laws were adopted that sought to balance three goals: preventing harm; protecting free expression and information access; and encouraging technical innovation and economic growth more generally.[5]

CDA 230 is by far the strongest safe harbor provision internationally because it immunizes online intermediaries unconditionally for the speech of others (outside of the area of intellectual property, sex trafficking, and federal criminal offenses). The strength of these protections against both liability and the considerable cost to platforms of even successful litigation is hard to overstate.[6] CDA 230 also immunizes online intermediaries for decisions to remove content, including lawful speech, from their services. CDA 230 thereby serves two parallel goals. First, to support the development of the internet ecosystem and freedom of expression by limiting risk and liability for relevant services acting as intermediaries. Second, to provide maximum space for such services to apply voluntary mechanisms to address potentially illegal and harmful content. Another important U.S. law, the Digital Millennium Copyright Act (DMCA), creates a detailed "notice-and-takedown" system for content alleged to infringe copyright.

In the EU, the ECD provides similar safe harbors for the activities of specific intermediaries, across a wider set of legal issues. The European safe harbors effectively create a notice-and-takedown system (or "notice and action," since an intermediary may respond in other ways besides taking content down) for content ranging from copyright infringement to hate speech. Rules vary from country to country, and are generally not spelled out in detailed statutes. Once an online intermediary obtains knowledge or awareness about illegal content, it loses immunity under the ECD and risks becoming liable. European courts have denied statutory immunities to intermediaries that were too "active" in engaging with user content – contrasting with CDA 230's encouragement of moderation and editorial control. Notably, the ECD leaves room for injunctions and duties of care at the national level with respect to unlawful content (including removal). Such injunctions are limited by Article 15, which prohibits national lawmakers from imposing general monitoring obligations on intermediaries for illegal content.

II. The Same Principles in Changed Circumstances

Clearly, the core principles underlying these frameworks, including the tackling of harm, protection of freedom of expression, and support for innovation, remain valid today. Still, the environment has

changed significantly since the adoption of these laws. First, intermediary service providers such as Google Search, YouTube, and Facebook may be considered dominant players and have been found to be unprepared to tackle emergent phenomena. This, and the broader political "techlash" we are witnessing today, undercuts one of the political rationales underlying the safe harbors: minimizing business risks to emerging companies and technologies. The safe harbors are now often portrayed as sweetheart deals for already dominant companies. The legal certainty they provide for (potential) new entrants and smaller service providers, however, remains important as a source of support for innovation and competition. Overall, it's easy to critique existing laws, but replacing them with something better will require considerable attention to the real mechanics of intermediary liability law, and to the doctrinal choices discussed in this essay.

There is a clear need to guard against oversimplifications in the discussion. For instance, simply scrapping the current protection from the law will not provide a clean slate for the determination of intermediary liability. It will leave internet users and online expression subject to a complicated, fragmented, and uncertain mix of traditional legal doctrines. While years of litigation might ultimately lead to reasonable and workable rules, the harm to both innovation and internet users' rights in the interim would be significant. Any regime that imposes liability on speech intermediaries should comply with constitutional and human rights safeguards. Intermediary liability laws' restrictions on core democratic freedoms such as freedom of communication, speech, and association, as well as the right to privacy, must be necessary, proportionate, and provided for by law.

The second important shift in the environment for intermediary liability laws involves platforms' role in society. Concerns about the impact of out-of-control online speech dynamics and challenges posed to our democracies abound. Generally, online platforms and associated technologies and practices have become catalysts in much wider economic, cultural and social change. But online platforms are also essential entities in the online ecosystem for freedom of expression, transforming the setting for the regulation of speech and harm into a triangle of platforms, users, and regulators. Considering these circumstances, a balanced answer to the questions about their proper roles and responsibilities with respect to societal impacts, including harmful ones, remains an essential legal and regulatory challenge.

In addressing this challenge, a warning is due with respect to a singular focus on the roles and responsibilities of online platforms and infrastructural services. Although online platforms are attractive targets for regulation, due both to their ability to exercise control as well as to market concentration, such regulation can have a number of clear downsides, such as privatized censorship. Intermediary liability frameworks should be assessed in light of their impact on both platforms and end users, and weighed against the option of laws targeting the primary speakers that are ultimately responsible for the publication of illegal and/or harmful content or activity.

As a result of the changes discussed above, thought leaders and policy makers on both sides of the Atlantic have started to question whether service providers, in particular large ones, should be expected and required to do more, and prevented from "hiding behind" first-generation internet regulations. In addition, political pressure is mounting on platforms to be more restrictive toward speech that is not necessarily illegal but is considered harmful, such as viral deception and certain forms of hate speech that are protected under the First Amendment to the U.S. Constitution, Article 10 of the European Convention on Human Rights, and Article 11 of the EU Charter of Fundamental Rights.

Thus, intermediary liability laws in the U.S. and Europe are a key issue for the TWG to consider and examine, as the debates about the possible revision of relevant statutory regimes are in need of robust guidance and well-informed recommendations. To provide input to the debates in Europe and the

United States and provide the basis for higher-level recommendations, the subsequent sections lay out the key components of intermediary liability laws, seen from a transatlantic perspective.

Specifically, we review the "dials and knobs" available to lawmakers seeking to update intermediary liability laws to account for present-day concerns. We break down key doctrines or provisions from existing law or current policy discussions into modular elements, focusing on their ramifications for free expression in particular. Of course, in real-world legislation, these elements rarely occur in isolation – and combining them can produce new effects. For example, a law holding platforms liable for deceptive speech they "know" about may mean something different depending on whether the law lets users explain and defend their posts. For the purposes of our discussion, however, isolating them can help in identifying options for well-designed intermediary liability laws.

III. Central Considerations

Platform liability laws affect internet users' free expression and access to information in two big ways. First, users' rights suffer if platforms are incentivized to "over-remove," taking down lawful speech in order to avoid liability or reduce costs. Over-removal in notice-and-takedown systems is well-documented.[7] Second, liability risks can deter innovators from building – or investors from funding – open speech platforms in the first place. As a result, strict liability standards for users' expression on platforms have generally been considered incompatible with freedom of expression. Legal and human rights literature identifies the following as particularly critical tools to mitigate threats to free expression:

No monitoring: Not requiring platforms to proactively filter or police user expression

Human rights literature includes strong warnings against making platforms proactively monitor, police, or filter their users' expression. Many intermediary liability laws expressly bar such requirements, though they have gained traction in recent European legislation such as the EU Copyright Directive for the Digital Single Market and some drafts of the Terrorist Content Regulation. One concern is that technical filters, which may range from simple hash-based systems for recognizing duplicate files to more sophisticated AI-based processes, are likely to over-remove because they are bad at recognizing context – like news reporting or parody – or accommodating changing legal interpretations. (Filtering, though, is relatively accepted for child sexual abuse images, which are unlawful in every context.) Another is that when platforms have to review and face over-removal incentives for every word users post, the number of unnecessary takedowns can be expected to rise. Under a law that requires monitoring, legal exposure and enforcement costs may also give platforms reason to allow only approved, pre-screened speakers, or to use Terms of Service (TOS) to prohibit controversial or legal gray-area speech. The liability risk and enforcement costs may also deter new market entrants from challenging incumbents.

Public due process: Using courts or other public authorities, not platforms, to decide what expression is illegal in most cases

As a protection for internet user expression rights, some countries reserve the responsibility for assessing certain claims against online content to courts or other government authorities. Platforms are immune until informed of the authority's legal determination that specific content is illegal. The government may also be subject to transparency obligations when it requires or suggests removal of content. This speech-protective standard typically has exceptions, requiring platforms to act of their own volition against highly recognizable and dangerous content such as child sexual abuse images. Lawmakers who want to move the dial toward harm prevention without having platforms adjudicate

questions of speech law can also create accelerated administrative or court processes or give platforms other responsibilities, such as educating users, developing streamlined tools, or providing information to authorities. Judicial review is particularly important and valuable for borderline cases involving disputed facts or nuanced legal doctrine.

Private due process: Requiring procedural protections for speakers when platforms take action against content

Building procedural protections into platforms' internal notice-and-takedown systems and terms of service enforcement can protect against over-removal. A widely supported civil society document, the Manila Principles, provides a menu of procedural protections with respect to notice and action. For example, a platform can be required or incentivized to notify the affected speaker, provide sufficient reasoning for its actions affecting her speech, and let her defend herself. The existence of such procedures may deter bad-faith notices in the first place. Claimants or accusers can also be required to include adequate information in notices and face penalties for bad-faith allegations. And platforms can be required to disclose raw or aggregate data about actions against content to facilitate public review and correction. Procedural protections for users affected by TOS enforcement – i.e., "private due process" – may also be required as a matter of consumer contract law. Self-regulation initiatives, which may have the partial aim to encourage reliance on TOS and prevent actual regulation from being passed, should come with robust private due process safeguards.

Public rule-setting: Regulating platforms' use of private Terms of Service enforcement

Platforms often take down disfavored but legal speech based on their TOS or Community Guidelines. To protect users' free expression rights and prevent undue bias in content moderation practices, a law might try to impose limitations on TOS enforcement against protected expression (possibly in combination with requiring the private due process discussed in the previous section). To ensure that governments do not fail in their own human rights obligations, they might also be prevented from relying on companies' TOS instead of publicly enacted law to regulate speech. However, limiting TOS enforcement would have some clear downsides from the perspective of tackling illegal and harmful activity. Rules designed to protect users' rights to expression, information, and due process may need to be balanced with Good Samaritan defenses (discussed below), which are designed to encourage platforms to moderate lawful but harmful content. TOS enforcement may also effectively serve to protect the free expression rights of vulnerable users. For example, reducing legal but abusive comments on platforms like Twitter can, as a practical matter, enable attacked or marginalized users to speak more freely. It can also make the platform attractive to other users, preserving its value as a forum for civil discourse. These arguments are particularly salient in countries like the U.S., where the law permits speech that violates widely held social norms or moral beliefs. Finally, in the U.S., the law also likely protects TOS enforcement as an exercise of platforms' own editorial rights. In Europe, too, freedom of expression and media pluralism warrant care in imposing community standards on platforms instead of allowing services to choose their standards freely within the boundaries of the law.

Remedies for speakers: Equal access to courts for speakers and victims of harmful content

Platform over-removal incentives come in part from asymmetry between the legal rights of accusers and those of speakers. Under most intermediary liability systems, including Europe's, victims of speech-based harms can sue platforms to get content taken down. Speakers, by contrast, can very rarely sue to get content reinstated or be compensated. (To our knowledge, such claims have succeeded only in Germany, Poland, and Brazil, and only very recently.[8]) This means that outside of

strict immunity regimes like the U.S.'s CDA 230, liability concerns consistently push toward removal. This asymmetry also distorts courts' opportunities to clarify the law, particularly when platforms enforce novel legal standards by allowing courts to review the claims of people seeking more content removal, but not the claims of people defending their expression rights. A few untested new laws in Europe, including the Audio Visual Media Services Directive and Copyright Directive, try to remedy this.[9] It is unclear how these new mechanisms will work in practice or how speakers' claims will intersect with platforms' power to take down speech using their TOS.

Consistent speech laws: Protecting the same expression online and offline

Content that might do only modest harm offline – like political disinformation spread by word-of-mouth to a few people – may do greater harm online, where it persists over time and can spread virally. Some lawmakers have responded to this concern by pressuring platforms to prohibit harmful-but-legal content voluntarily under their terms of service.[10] This approach reduces public rule making and due process for sensitive free expression issues. Others have proposed or enacted laws restricting online dissemination of speech that is otherwise legal.[11] This approach – which has long been strongly disfavored in human rights literature – resembles regulation of older media like broadcast or cable, which, for instance, have rules to protect minors. Applying such rules to internet platforms would put a greater burden on free expression rights, though, because it would affect everyday speech by internet users who rely on platforms to communicate.

Legal predictability: Bright-line rules versus fuzzy standards

Intermediary liability rules can hold platforms to flexible standards like "reasonableness" in responding to potentially unlawful user content, or prescribe specific steps. Both platforms and free expression advocates typically favor the latter because it increases predictability and reduces the role of platform judgment. Poorly calibrated process rules may encourage over-removal – if, for example, platforms automatically honor all takedown demands -- but this can be somewhat offset with private due process requirements, like counter-notice or transparency.

Private vs. public speech: Respecting communications privacy and targeting public (illegal and harmful) speech

Appropriate IL rules may be different for fully public communications (like a blog post or tweet) as compared with private communications (like email or a post to a small closed Facebook group). Existing legal frameworks for communications services include protections for communications privacy, with some of these protections also applying to new services (beyond traditional telecommunications). Internet services have also increasingly integrated technical protections, including end-to-end encryption in services such as WhatsApp, Signal, and Telegram. As internet users migrate toward these private platforms (potentially as a result of content moderation practices), there is increasing pressure on service providers to police private communications or even build encryption backdoors. IL laws targeting private communications services should only do so while respecting communications privacy and security. The distinction between private and public speech is not a sharp one, and appropriate rules may vary depending on the type of illegal or harmful content. (In some cases, disseminating content may be legal in a private communication but not in a public one.)

Cross-border dimension: Respecting the global nature of internet speech and platforms

Across jurisdictions, IL laws or underlying law on issues such as hate speech or disinformation may set different standards that can create cross-border conflicts since global platforms are subject to legal pressure from around the world. For instance, one country may set a Good Samaritan defense

permitting platforms to remove lawful but harmful speech, while another country may limit a company's freedom to do so. Or a court in one jurisdiction may order the global removal of speech that is legal elsewhere. Intermediary liability laws should be respectful of the global nature of internet speech platforms and minimize cross-border conflicts limiting freedom of expression. Jurisdictions that want to increase enforcement of their local laws by tightening IL standards (like NetzDG) generally should not require the entire worldwide platform to operate according to their standards.

IV. Standards for Platforms' General Conduct

Some recently proposed standards focus on platforms' responsibility in their overall operations, rather than on case-by-case liability for individual items of unlawful content.

"Due diligence" or "duty of care" standards

At their core, IL laws are concerned with the question of what liability exists for *specific instances* of illegal content or activity. IL laws have established that such liability should be limited in nature (e.g., only after the service has actual knowledge) and not strict (liability imposed regardless of knowledge, establishing de facto proactive duties to monitor to prevent liability). Considering the current direction of discussions about IL, the question is whether and what policy options exist between those conditions. Some recent European laws and proposals move away from penalizing platforms for individual incorrect decisions about specific user expression, and instead seek to regulate platforms' overall content management operations and create new forms of administrative oversight with respect to these frameworks.[12] This approach might crudely be analogized to food safety standards that accept a certain number of insect or other contaminant parts-per-million, on the basis that requiring a smaller margin of error would impose disproportionate costs on both the regulated entity and society. For instance, under this model, a platform that meets a "duty of care" or "due diligence" standard in its overall content moderation system would not be punished for one-off mistakes. One can also imagine more specific targets for content moderation practices. In countries willing to accept significant regulatory review and standard setting for platforms, these approaches may represent an important new way forward. They could build on existing approaches with respect to risk management and fundamental rights impact assessment, requiring platforms to consider risks to freedom of expression, due process, non-discrimination and minority participation in public discourse.

Transparency requirements

Transparency reporting has emerged as a practice to create more accountability for removal of content by platforms. Industry transparency reporting practices have developed over the last decade, with reports providing insights into the number of requests to take down allegedly unlawful content in different categories. Transparency reporting is required in some new intermediary liability laws, such as Germany's NetzDG and the EU's proposal for a Terrorist Content Regulation, and co-regulatory frameworks for hate speech and disinformation. The reports have become important sources of evidence, but it remains difficult to compare different platforms' reports because of differences in their reporting procedures and standards.

V. Liability Based on Knowledge or Control

Traditional tort doctrines typically held publishers or distributors liable based on their editorial control or knowledge about unlawful content. Similar standards appear in many IL laws, although platforms

often differ from pre-internet publishers or distributors in the volume of third-party expression they handle and in their relatively weak incentives to defend it.

Knowledge and other "mental state" standards

Many legal systems hold platforms liable for continuing to host or transmit illegal content once they "know" or "should know" about it. This is for instance the case under the European intermediary liability framework, Article 14 of the e-Commerce Directive specifically. Similar standards exist in U.S. criminal law and copyright law.[13] Others reject this standard, considering it too likely to incentivize over-removal. Laws that use knowledge standards can reduce this problem somewhat by defining "knowledge" narrowly or adding elements like private due process.

Controlling or "active" platform standards

Most IL laws strip immunity from platforms that are too actively involved in user content, e.g., because they help create or solicit particular material, or optimize, select and/or promote it in a commercial context or for profit-making purposes.[14] Some version of this rule is necessary to distinguish platforms from content creators. But laws that reward passivity also generate legal uncertainty – and may deter innovation or lead to over-removal – as platforms consider new features that go beyond bare-bones hosting or transmission. They also risk deterring platforms from moderating content at all, for fear of losing immunities.

"Good Samaritan" rules to encourage moderation

Platforms that want (for economic or other reasons) to weed out illegal or offensive content may be deterred by both "knowledge" and "control" liability standards. Plaintiffs can use platforms' moderation efforts as evidence of editorial control, or argue that the platform knew about content that a moderator saw but did not take down. This concern underlies the broad immunities the U.S. established in CDA 230. The current European framework lacks a Good Samaritan defense and platforms also risk losing their safe harbor protection if they more proactively address illegal and harmful content. This has complicated the development of self- and co-regulation to tackle illegal and harmful content online.

VI. Using Different Rules for Different Problems

Real-world laws typically combine elements listed here, which allows lawmakers to more carefully calibrate trade-offs affecting free expression. The potential downside is that complex laws can increase operational costs for platforms, potentially leading them to simplify by being too restrictive.

Variations based on legal claim

IL laws often require special or more urgent treatment for particularly harmful or highly recognizable content, such as child sexual abuse material. By contrast, they may provide stronger free expression protections for claims that platforms cannot reasonably assess because of nuanced legal doctrines or disputed facts, such as in the case of defamation.

Variations based on a platform's technical function and relation to user expression

Many IL laws put the risk of liability on the entities most capable of carrying out targeted removals – like taking down a single comment instead of a whole page or website. This is also consistent with the internet's "end to end" technical design principles. Thus, infrastructure providers like ISPs or domain registries generally have stronger legal immunities than consumer-facing platforms like YouTube.

Many recently proposed IL laws, like the 2018 amendments to CDA 230 in the U.S., have not reflected this principle.[15]

Variations based on a platform's size

Recently, experts have raised the possibility of special obligations for mega-platforms like Google or Facebook. Drafting such provisions without distorting market incentives, driving bad actors to less strictly policed, smaller platforms or punishing unusual platforms like Wikipedia would be challenging. In principle, though, it might improve protections on the most popular forums for online expression without imposing such onerous requirements that smaller market entrants couldn't compete.

VII. Judicial Actions Against Platforms

Platforms' actions against user-generated content can be shaped both by *direct* legal mandates (like injunctions) or the *indirect* influence of potential future claims. In deciding whether to remove legal gray-area content like crude parodies, for example, platforms may act based on their expectations or fears of what a court *might* do if the content stays up and a plaintiff or prosecutor brings a legal claim. Thoughtfully tailoring the availability of financial damages or injunctive relief in the platform context can help protect lawful expression.

Cost to platforms

Over-removal incentives (or incentives to stop offering services completely) are likely to be greatest when platforms fear high damages, regulatory attention that can lead to other costs or business impact, or business-altering injunctions (like having to turn off popular features).

Scope of injunctions

Because countries vary in their laws regarding expression, a platform takedown order issued in one country can affect speech and information that is legal in another. Geographically limited orders can mitigate this problem, but mean that harms may be addressed less effectively. Courts can also issue time-limited orders, allowing content to become accessible again after a certain period.

Options other than taking down or reinstating content

Increasingly, content-related issues in online platforms are dealt with through measures short of simply removing the material. For example, controversial but lawful content may be demoted in rankings, demonetized, or delisted in search results. IL laws can replace binary take-down/leave-up outcomes by stipulating more tailored remedies. For example, platforms can show users a warning before viewing certain content, cut it off from ad revenue, or show it in response to some search queries but not others. In principle, IL law could also regulate the algorithms that platforms use to rank, recommend, or otherwise amplify or suppress user content. Such a law, however, would be complex to define, enforce, and administer.

VIII. Conclusions

There are several structural reasons for revisiting intermediary liability laws that were adopted in the 1990s. When doing so, lawmakers should continue to be informed by the principles underlying these laws, including freedom of expression, as well as more than two decades of experience in providing for intermediary liability provisions and associated policies. Simply scrapping existing safe harbor provisions currently in place would in no way resolve many of the issues outlined here, and would

inevitably cause significant legal uncertainty and harm to competition and fundamental rights of internet users. In view of this, this discussion paper offers a systematic overview of key elements to consider in potential revisions and design of new IL laws and ways in which these elements can be approached in a balanced manner.

Lawmakers considering potential revisions to IL laws should craft any amendments carefully to avoid incentivizing platforms to act against the rights and interests of their users. The concerns and doctrinal tools listed under Central Considerations in section III are particularly key for this purpose, and should serve as guiding parameters. For example, under laws requiring platforms to remove unlawful content, "private due process" protections such as notice and appeals for the affected users can serve to protect expression and information rights. Lawmakers can further refine IL laws using transparency requirements, legal obligations tailored to intermediaries' technical functions, and other doctrinal tools discussed in sections IV-VII. By adjusting the "knobs and dials" set forth in this discussion paper, lawmakers can strike an appropriate and proportionate balance between reducing online harms, protecting fundamental rights, and promoting innovation and competition.

Notes

[1] Professor of Law, appointed to the Chair "Fundamental Rights and the Digital Transformation," established at the Interdisciplinary Research Group on Law Science Technology & Society (LSTS), Vrije Universiteit Brussel (VUB), with the support of Microsoft; Senior Researcher, Institute for Information Law (IViR), Faculty of Law, University of Amsterdam.

[2] Director of Intermediary Liability, Stanford Center for Internet and Society (CIS); former Associate General Counsel, Google. Stanford CIS funding information is available at http://cyberlaw.stanford.edu/about-us.

[3] An earlier version of this TWG discussion paper was prepared for the Santa Monica meeting of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression in May 2019. Sections III-VII are adapted and expanded from Daphne Keller, *Build Your Own Intermediary Liability Law: A Kit for Policy Wonks of All Ages* (June 2019), https://balkin.blogspot.com/2019/06/build-your-own-intermediary-liability.html?m=1. We would like to thank Sherwin Siy and Jan Gerlach of the Wikimedia Foundation and João Pedro Quintais of the University of Amsterdam for their input in drafting the initial document. Results of the discussions at the Santa Monica meeting have been incorporated into the text.

[4] The distinction between *illegal* content or activity and *harmful* content or activity is important. Harmful content and behavior includes speech and activity that the law permits but that still cause harm.

[5] Intermediary liability laws can also raise questions regarding other fundamental rights (including due process, communications freedom and confidentiality, data privacy, freedom of association and non-discrimination) and policy goals (e.g., competition, trade policy).

[6] Liability in the narrow sense can be about civil liability for damages caused by or criminal liability for third-party content and activity. In addition, there is an important question about the possibility to impose injunctions, for instance in a possible situation of negligence.

[7] Listing of empirical studies documenting platform removal of lawful speech under notice-and-takedown systems (last updated 2018): http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

[8] There may be many more cases in which an initial threat of a lawsuit by a speaker affected by a removal gets resolved by the platform reinstating the content.

[9] The AVMSD does so by giving users recourse to administrative review when platforms remove allegedly unlawful content.

[10] In Europe, for instance, lawmakers are discussing referral mechanisms in combination with calls on service providers to ban terrorism content through their terms of service. For a discussion, see Van Hoboken, "The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications," TWG discussion paper, Institute for Information Law, Amsterdam, 3 May 2019. Available at https://www.ivir.nl/publicaties/download/TERREG_FoE-ANALYSIS.pdf.

[11] See for instance the UK's Online Harms White Paper. Available at https://www.gov.uk/government/consultations/online-harms-white-paper.

[12] See in particular the NetzDG law in Germany. For a discussion of NetzDG, see Tworek and Leerssen, "An Analysis of Germany's NetzDG Law," TWG discussion paper, 15 April 2019. Available at https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

[13] 18 U.S.C. 2252, 2258A, 2258B (knowledge-based liability and obligations for intermediaries regarding child sexual abuse material); 17 U.S.C. 512 (intermediaries lose DMCA immunity based on actual or "red flag" knowledge).

[14] This is the case for the e-Commerce Directive in the European Union. U.S. intermediaries lose CDA 230 immunity if they materially contribute to the unlawfulness of content, *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC,* 521 F. 3d 1157, 1168 (9th Cir. 2008), and lose DMCA immunity based on right and ability to control infringing content from which they directly benefit. 17 U.S.C. 512.

[15] The Allow States and Victims to Fight Online Sex Trafficking Act, commonly known as FOSTA, which was enacted in 2018, draws no distinctions between obligations of infrastructure providers and those of edge-of-network, user-facing platforms. H.R. 1865, 115th Cong. (2018).

TRANSATLANTIC WORKING GROUP

An Examination of the Algorithmic Accountability Act of 2019†

Mark MacCarthy, Georgetown University[1]

October 24, 2019

Contents

Introduction

The Algorithmic Accountability Act of 2019, sponsored by Senators Cory Booker (D-NJ) and Ron Wyden (D-OR), with a House equivalent sponsored by Rep. Yvette Clarke (D-NY), requires companies to assess their automatic decision systems for risks to "privacy and security of personal information" and risks of "inaccurate, unfair, biased, or discriminatory decisions." They must also "reasonably address" the results of their assessments. The bill empowers the Federal Trade Commission (FTC) to resolve by regulation the crucial details of these requirements. It is not likely to pass Congress on its own, but it might become part of a new national privacy law currently under consideration in Congress. If passed, it would apply to the AI systems that platforms increasingly deploy to detect and counter hate speech, terrorist material and disinformation campaigns and would require the platforms to conduct fairness assessments of these AI systems and fix issues of bias uncovered in these studies. Even without a legislative mandate, however, platforms should rigorously review algorithmic content moderation systems for fairness and accuracy and should establish and

† One in a series: A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

maintain effective, easy-to-use complaint mechanisms whereby people wrongly labeled as purveyors of harmful material can obtain redress.

Summary of the Bill

On April 10, 2019, Senators Booker and Wyden introduced S. 1108, the Algorithmic Accountability Act of 2019.[2] Rep. Clarke introduced an identical companion bill, H.R. 2231, in the House.[3] The Senate bill was referred to the Senate Commerce Committee while the House bill went to the House Energy and Commerce Committee, the committees that will deal with privacy legislation later this year.

The bill contains an exemption for small businesses. It would apply to companies under the FTC's jurisdiction that make more than $50 million per year in gross receipts or have data for at least 1 million people or devices. It would also apply to "data brokers" (a company that "collects, assembles, or maintains personal information concerning an individual who is not a customer or an employee of that entity in order to sell or trade the information or provide third-party access to the information") regardless of their revenue or the number of people whose data they hold.

The bill would direct the Federal Trade Commission to pass regulations within two years to require these companies to conduct studies of their high-risk automated decision systems "for impacts on accuracy, fairness, bias, discrimination, privacy, and security." Companies must conduct an assessment for new systems "prior to implementation" and for existing systems "as frequently as the Commission (FTC) thinks is necessary." These impact assessments may be made public, but need not be, at the "sole discretion" of the company. They must conduct these assessments "if reasonably possible, in consultation with external third parties, including independent auditors and independent technology experts" and must "reasonably address in a timely manner the results of the impact assessments."

An automated decision system is a "computational process, including one derived from machine learning, statistics, or other data processing or artificial intelligence techniques, that makes a decision or facilitates human decision making, that impacts consumers."

Companies must conduct assessments only for their "high-risk" automated decision systems. An automated decision system is high-risk if it satisfies *any* of the following conditions:

- It poses "a significant risk to the privacy or security of personal information ... or of resulting in or contributing to inaccurate, unfair, biased, or discriminatory decisions impacting consumers."

- It "makes decisions, or facilitates human decision making" that "alter legal rights of consumers … or otherwise significantly impact consumers" when these decisions are based on "attempts to analyze or predict sensitive aspects of their lives, such as their work performance, economic situation, health, personal preferences, interests, behavior, location, or movements."

- It "involves the personal information of a significant number of consumers regarding race, color, national origin, political opinions, religion, trade union membership, genetic data, biometric data, health, gender, gender identity, sexuality, sexual orientation, criminal convictions, or arrests."

- It "systematically monitors a large, publicly accessible physical place."
- It "meets any other criteria established by the Commission in regulations…"

The bill also requires companies to conduct a "data protection impact assessment" for "high risk information systems." An information system is a database that "involves personal information, such as the collection, recording, organization, structuring, storage, alteration, retrieval, consultation, use, sharing, disclosure, dissemination, combination, restriction, erasure, or destruction of personal information." An information system is high risk when it poses significant risks to the privacy or security of personal information, or as above when it collects sensitive information, monitors a large, public space, or meets other criteria established by the commission. A data protection impact assessment is narrower than an automated data system impact assessment, focusing only on "the extent to which an information system protects the privacy and security of personal information the system processes."

The bill's basic requirement to conduct impact assessments systems appears to be modeled on several provisions of the European General Data Protection Regulation (GDPR). Article 22 of the GDPR provides for a right for a data subject "not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[4] Article 35 requires companies to conduct "data impact assessments" of the risks of data processing operations "to the rights and freedoms of natural persons" and their effect "on the protection of personal data."[5]

The FTC is provided substantial authority to enforce the provisions of the bill, treating violations as if they were violations of rules defining "unfair and deceptive practices" under its authorizing statute. The bill does not allow the FTC to exceed its current authority in drafting implementing rules or its enforcement actions.

Despite much discussion of the need for companies to disclose the source code or the formula of their algorithms and to provide explanations of machine learning algorithms,[6] the bill makes no such demand on companies. The bill requires companies to assess their algorithms for conformity to various standards such as privacy, security, and fairness. But under the proposed bill, they can keep their formulas and source code secret, and they need not provide explanations of how they work. It would not be too hard, however, to add transparency and explainability requirements to the bill if it began to move through the congressional process.

Interpretation of the Bill

Under the proposed bill, the FTC has broad authority to define the key terms in the requirement to conduct impact assessments – discrimination, bias, unfairness, privacy and security. It is likely that the FTC would think of the requirement to avoid unfairness in terms of its current authority to prevent companies from engaging in an unfair act or practice, that is, an act or practice that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition."[7]

The FTC has brought cases under its unfairness authority in connection with several types of informational injury including financial harm, physical injury, reputational harm and unwanted intrusion.[8] Companies conducting algorithmic assessments covered by the FTC's rules developed under the new law would likely have to include whether the use of the algorithms in company decisions would cause or be likely to cause harms of this nature.

The mandate for companies to conduct analyses for bias and discrimination is different. It would be new to the FTC, but it is not new for companies that already need to comply with a range of U.S. laws forbidding discrimination against protected classes in a variety of contexts. The current anti-discrimination laws prohibit discrimination against protected groups in employment,[9] credit,[10] housing,[11] use of genetic information,[12] and health care and health insurance.[13]

The use of disparate impact analyses in anti-discrimination law is complex and controversial.[14] But companies often conduct these analyses to assess and control legal risk of non-compliance with current anti-discrimination laws. In general a disparate impact assessment analysis has three stages: evidence of a disproportionate impact caused by a policy or procedure; assessment of whether and to what extent the policy or procedure serves a valid purpose; and assessment of whether there are alternative policies or procedures that would achieve the legitimate objective with a less disparate impact.

For instance, a potential legal risk in the employment context would be the presence of a disproportionate adverse effect on a protected class that takes place unintentionally through a policy, procedure or decision algorithm that does not base a hiring or promotion decision explicitly on protected class status. Potential violations of statistical parity in employment contexts can be detected by a company study assessing compliance with the 80% rule of thumb, which suggests, for instance, that if a company hires 10% of white applicants then it should hire no less than 8% of African American applicants.[15]

Companies frequently conduct disparate impact analyses to determine if their decisions depart from statistical parity among protected groups, and if so, whether there are legitimate reasons for this departure and alternative decisions procedures that would create less impact. This is what happened with Amazon's attempt to develop a hiring algorithm for software engineers. It experimented with using historical data to train an automated method of selecting promising applicants, did disparate impact analyses, found that these algorithmic results disproportionately rejected women, and fixed the factors that were leading to this disparate impact but encountered others it could not fix. Aware of the legal risk under current laws that bar discrimination on the basis of sex, the company wisely used the conclusions of its disparate impact analysis to abandon that attempt at recruitment automation. [16]

The FTC is likely to interpret the mandate in the proposed law for companies to review an algorithm for bias and discrimination as similar to the way in which companies currently conduct disparate impact analysis under existing anti-discrimination law. This would mean an extension of the range of areas in which companies might be required to conduct these analyses beyond what is already required under current law.

It is possible that the law is intended to remedy a purported exemption from these laws for algorithms. The press statement accompanying the introduction of the bill suggests this, saying, "Algorithms

shouldn't have an exemption from our anti-discrimination laws."[17] But algorithms do not have such an exemption. Current anti-discrimination laws cover the use of old-fashioned statistical methods such as regression analysis to assess people, as well as the most up-to-date machine learning techniques.[18]

The press release also cites action brought against Facebook for discrimination in housing ads. But that action was brought under the Fair Housing Act and clearly demonstrates that the Fair Housing Act covers discriminatory housing ads that use targeting algorithms. In announcing the action against Facebook, Ben Carson, Secretary of Housing and Urban Development, said, "Using a computer to limit a person's housing choices can be just as discriminatory as slamming a door in someone's face."[19]

So it is most likely that the bill is aimed at uses of algorithms that are close to the line in terms of current discrimination law, or are not covered at all under current law.[20] Is it illegal for a financial institution to tailor its online website so that the best credit card offers appear to people who seem to be the best credit risks? Is it a violation of employment discrimination law when women are less likely than men to be shown ads for high-paying jobs? Is a facial recognition program that is less accurate for blacks than for whites discriminatory? The law is less than clear in these areas, and the proposed bill might best be interpreted as a directive to the Federal Trade Commission to provide clarity.

But such clarifications are not in the bill itself. The proposal does not specify that any particular uses of algorithms are discriminatory. Its definitions are broad enough to include such common uses of algorithms as ad targeting, facial recognition, search engines, fraud detection systems, and identity verification systems. But its key terms of bias, discrimination and unfairness are left up to the FTC to define.

There are other matters that the FTC will have to clarify in implementing regulations. For instance, what will define whether consultation with external third parties in the conduct of assessments is "reasonably possible?" Will the FTC define the circumstances or leave it up to the companies?

Probably most important is the unresolved question of what companies might have to do to "reasonably address … the results of the impact assessments." In the case of an ad for high-paying jobs that target men, the agency could conclude that the only way to "reasonably address" the harm of employment ads targeted at men would be to alter them so that they preserved statistical parity for men and women – the chances of a man getting the ad have to be approximately the same as the chances of a woman getting the same ad. This would be a substantial intrusion into the ad marketplace, but it seems to be clearly within the authority granted to the FTC under the proposed legislation.

Effect on Content Moderation Programs

The bill does not directly affect content moderation programs and their content rules regarding hate speech, terrorist material and disinformation campaigns. But platforms increasingly rely on algorithms to identify and counter harmful speech. The proposed bill could indirectly affect the content moderation programs established by platforms through its requirement to conduct algorithmic assessments of these content moderation algorithms and its further authority to require unspecified fixes based on the results of these assessments.

These content moderation algorithms clearly fit the bill's definition of high-risk automated decision systems. These algorithms might make or contribute to a decision that certain people are likely to be part of a hate group, disinformation campaign or terrorist plot through a computerized analysis of their "interests, behavior, location or movements." Such a determination could certainly "significantly impact" these people, for instance, through barring them from access to the platform, and might even alter their "legal rights" if their identifying information was turned over to law enforcement. A platform's automated methods of detecting harmful content might also affect the privacy of platform users by collecting large amounts of information and making inferences to sensitive attributes.

FTC regulations under the new law are likely to require assessments of the accuracy, fairness and disparate impact of content moderation algorithms. They will not require disclosure of the source code or the formula, or a requirement for explanation of the operation of these algorithms. Under the law, these assessments can be made public only with the consent of the platforms conducting them, but it is highly likely that the FTC will require access to the results of the studies and maybe even to the studies themselves.

It is not clear what the FTC will require in terms of the results of these assessments. They will at minimum examine whether the data gathered as part of the development of these algorithms is consistent with FTC privacy guidelines and conformity to the requirements of the new privacy law Congress is considering, which would likely provide FTC with new enforcement authority.

The commission will also likely examine whether a particular use of an algorithm in content moderation decisions is an unfair practice and whether it causes or is likely to cause substantial injury to consumers that they cannot reasonably avoid and has no compensating benefit. In doing this, it is likely to consider the various informational injuries − financial harm, physical injury, reputational harm and unwanted intrusion − that it has considered in past cases.

The commission will examine whether the platforms' assessments of their content moderation algorithms reveal actionable disparate impact under new discrimination standards that the commission has developed. For instance, platforms using algorithms that disproportionately identify users from a protected class as more likely to post material violating platform content rules might need to engage in special scrutiny of these algorithms and justify their use both in terms of their accurate contribution to content moderation goals and the demonstrated lack of alternative algorithms that achieve that purpose with less impact on protected classes.

It might be possible for the FTC to examine the political neutrality of a content moderation program under the requirement for algorithmic fairness. For instance, the FTC might require platforms to examine their measures to address political disinformation campaigns to see whether they affected campaigns of Democrats and Republicans alike, and under its new regulations it might require demonstrating that any political disparate impact was the smallest consistent with an effective program to counter disinformation campaigns.

However, this is so far from the FTC's core mission that it is unlikely to do so under the bill as it is currently written. But a political neutrality requirement could be added in further consideration of the bill.

Political Assessment

In the one public reaction to the bill's introduction, law scholars Margot Kaminski and Andrew Selbst praised the bill, but suggested several improvements.[21] They suggest that the bill provide for additional enforcement and a clearer statement that algorithmic bias is illegal. They also argue that the bill needs greater public input and providing public audits only when "reasonably possible" is too narrow. Third, they suggest greater transparency in the assessments. They recognize that full transparency might jeopardize trade secrets and allow hostile actors to game the system and suggest an FTC report as one way to provide more openness while still allowing for more transparency.

These suggested improvements might be the enemy of the good. The chances that Congress will act on the bill as introduced are slim. The authors are all Democrats in a Senate controlled by Republicans. In the absence of a single Republican cosponsor, Republican committee chairs are unlikely to hold hearings on the bill or schedule it for a markup. Republican Senate leadership would not be likely to provide floor time for the bill without strong support by senior Republican senators. In the House, the lone sponsor is the Vice Chair of the House Energy and Commerce Committee, to which her bill was referred for action. She is thus well-positioned to encourage subcommittee and full committee chairs to hold hearings and markups on her bill. She cannot, however, schedule this on her own initiative.

As a result, the bill is unlikely to move on its own. It might however be an element of a larger privacy bill that is currently under consideration in both the House and Senate. The bill's focus on the bias, discrimination and unfairness in automated decision systems is mirrored in a draft bill proposed by Intel.[22] In addition, draft legislation from the privacy advocacy group the Center for Democracy and Technology instructs the FTC "to define and prohibit unfair targeted advertising practices," a specific use of an automated decision system that is also addressed in the algorithmic accountability proposal.[23] Finally, Congress is looking for guidance to Europe's GDPR, which, as previously noted, has a similar requirement as this bill with regard to automated decision systems.

Prospects are better in the House than in the Senate for attaching the bill to privacy legislation. The House bill has 26 cosponsors including Rep. Karen Bass (D-CA), the Chair of the Congressional Black Caucus, and a substantial number of caucus members have shown concern over algorithmic bias and the inadequacies of current law in dealing with it. This group has the political clout in the House to block movement on privacy legislation unless it includes measures addressing algorithmic bias.

In sum, Congress is considering the issues raised by the bill in regard to the fairness, accuracy and privacy of automated decision systems as it contemplates passage of a new national privacy law. The bill's most likely path to becoming law is by having significant elements of it become part of this larger privacy law. As mentioned earlier, additions to the bill requiring explanations, transparency and political neutrality might emerge during the political horse-trading that characterizes the movement of any complex legislation through Congress.

Conclusion and Recommendations

Platforms are properly focused on developing algorithms that can help them identify and counter harmful material on their systems. This is especially important for hate speech, disinformation

campaigns and terrorist material that can have harmful effects on other platforms and in the real world.

The Algorithmic Accountability Act of 2019 should remind platforms, however, that decisions to remove content and to take action against people who post content in violation of their terms of service can pose significant risks of harm to people if they are based on inaccurate or unfair automated systems. As a result of content moderation decisions, users can be barred from platforms; their names might be entered into shared databases that could form the basis for widespread denial of platform services; and their personal information might be turned over to law enforcement or security officials for prosecution or further surveillance. If these actions are taken in error more often for users in a protected class, for instance, this might exacerbate already substantial disparate impacts experienced by members of vulnerable groups.

One recommendation that emerges from this consideration of the bill is that even if it does not become law, systematic assessments of the AI systems used to detect harmful content – to ensure the greatest possible fairness and accuracy – would be a valuable addition to platform content moderation programs. A further recommendation would be to assess the tradeoffs platforms have to make between taking down harmful material in error and leaving up harmful material in error, taking into account the damage that can be done to a person's reputation and opportunities by wrongfully designating that person as a member of a hate group, terrorist organization or disinformation campaign. Because such mistakes will inevitably occur in complex systems like social media platforms, there need to be robust redress mechanisms through which people wrongfully labeled as purveyors of harmful material can clear their names.

Notes

[1] Mark MacCarthy is adjunct professor at Georgetown University, where he teaches courses in the Graduate School's Communication, Culture, and Technology Program and in the Philosophy Department. He is also Senior Fellow at the Institute for Technology Law and Policy at Georgetown Law, Senior Policy Fellow at the Center for Business and Public Policy at Georgetown's McDonough School of Business and Senior Fellow with the Future of Privacy Forum.

[2] S. 1108 – 116th Congress: Algorithmic Accountability Act of 2019, https://www.congress.gov/bill/116th-congress/senate-bill/1108.

[3] H.R. 2231 – 116th Congress: Algorithmic Accountability Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231?q=%7B%22search%22%3A%5B%22yvette+clarke%22%5D%7D&s=5&r=12.

[4] See Article 22 of GDPR, available at https://gdpr-info.eu/art-22-gdpr/.

[5] See Article 35 of GDPR, available at https://gdpr-info.eu/art-35-gdpr/. A data impact study is "particularly required" in the case of "automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person" or sensitive information, or monitoring of a large public area. See also Article 29 Working Group, Guidelines on Data Protection Impact Assessment, October 13, 2017, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236. I'm grateful to Joris van Hoboken for pointing out this connection to Article 35.

[6] For instance, Articles 13–15 of GDPR provide rights to "meaningful information about the logic involved" in automated decisions.

[7] 15 U.S.C. § 45(n). Section 3(d)(1) of the bill supports this interpretation by mandating the Commission to treat a failure to conduct an impact assessment required under the Commission's rule as "as a violation of a rule defining an unfair or deceptive act or practice" under the Federal Trade Commission Act.

[8] Federal Trade Commission, Staff Comments to NTIA on Consumer Privacy, November 13, 2018, pp. 8-9, https://www.ftc.gov/system/files/documents/advocacy_documents/ftc-staff-comment-ntia-developing-administrations-approach-consumer-privacy/p195400_ftc_comment_to_ntia_112018.pdf.

[9] Title VII of the Civil Rights Act of 1964 makes it unlawful for employers and employment agencies to discriminate against an applicant or an employee because of such individual's "race color, religion, sex, or national origin." It is enforced by the Equal Employment Opportunity Commission and state fair employment practices agencies. See 42 U.S.C. §2000e-2 available at http://www.law.cornell.edu/uscode/text/42/2000e-2.

[10] The Equal Credit Opportunity Act makes it unlawful for any creditor to discriminate against any applicant for credit on the basis of "race, color, religion, national origin, sex or marital status, or age 15 U.S.C. § 1691 available at http://www.law.cornell.edu/uscode/text/15/1691. The Federal Reserve Board originally enforced the Equal Credit Opportunity Act, but the Dodd-Frank Act of 2011 transferred jurisdiction to CFPB. See Consumer Financial Protection Bureau, CFPB Consumer Protection Laws: ECOA, June 2013, p. 1 available at: https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf.

[11] Title VIII of the Civil Rights Act of 1968, the Fair Housing Act, prohibits discrimination in the sale, rental or financing of housing "because of race, color, religion, sex, familial status, or national origin." The act also protects people with disabilities and families with children. It is enforced by the Department of Housing and Urban Development. 42 U.S.C. 3604 available at http://www.law.cornell.edu/uscode/text/42/3604.

[12] The Genetic Information Nondiscrimination Act of 2008 prohibits U.S. health insurance companies and employers from discriminating on the basis of information derived from genetic tests. Pub. L. No. 110-233, 122 Stat. 881 available at https://www.govinfo.gov/content/pkg/PLAW-110publ233/pdf/PLAW-110publ233.pdf.

Enforcement is divided among a number of agencies including the Department of Health and Human Services (for health insurance) and the Equal Employment Opportunity Commission (for employment).

[13] Section 1557 of the Affordable Care Act of 2010 prohibits discrimination in health care and health insurance based on race, color, national origin, age, disability, or sex. 42 U.S.C. § 18116, available at https://www.law.cornell.edu/uscode/text/42/18116.

[14] For a discussion, see Software & Information Industry Association, Algorithmic Fairness, 2017, pp. 8-12.

[15] "A selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact…" Uniform Guidelines on Employee Selection Procedures (1978), 29 C.F.R. § 1607.40 (1987), available at http://uniformguidelines.com/uniguideprint.html.

[16] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, October 9, 2018 at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[17] Wyden, Booker, "Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms," April 10, 2019 at https://www.wyden.senate.gov/news/press-releases/wyden-booker-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms-.

[18] The Obama Administration recognized this when they recommended that regulatory agencies "should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law in such cases." See Executive Office of the President, "Big Data: Seizing Opportunities, Preserving Values," May 2014, p. 60 available at https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf. The Consumer Financial Protection Bureau claimed jurisdiction over the company Upstart that used alternative data and algorithms to assess lending decisions with respect to compliance with the Equal Credit Opportunity Act. See Consumer Financial Protection Bureau, CFPB Announces First No-Action Letter to Upstart Network Company to Regularly Report Lending and Compliance Information to the Bureau, September 14, 2017, https://www.consumerfinance.gov/about-us/newsroom/cfpb-announces-first-no-action-letter-upstart-network/.

[19] Katie Benner, Glenn Thrush and Mike Isaac, "Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says" New York Times, March 28, 2019, at https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html.

[20] The press release says the bill targets discrimination such as the "houses that you never know are for sale, job opportunities that never present themselves, and financing that you never become aware of – all due to biased algorithms."

[21] Margot E. Kaminski and Andrew D. Selbst, "The Legislation That Targets the Racist Impacts of Tech; A proposed law would make big companies determine whether their algorithms discriminate, but it's lacking in some big ways," New York Times, May 7, 2019, https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html.

[22] Section 4 of the draft bill requires companies to conduct a study to ensure that their automated decision processes are "reasonably free of bias and error," https://usprivacybill.intel.com/wp-content/uploads/IntelPrivacyBill-01-28-19.pdf.

[23] Section 6 of the draft bill contains this instruction to the FTC, https://cdt.org/files/2018/12/2018-12-12-CDT-Privacy-Discussion-Draft-Final.pdf.

TRANSATLANTIC WORKING GROUP

RESEARCH BRIEF

U.S. Initiatives to Counter Harmful Speech and Disinformation on Social Media†

Adrian Shahbaz, Freedom House

June 11, 2019

Harmful speech

There are limited, if any, legislative efforts in the United States that directly target hate speech. Attempts to combat hate speech through legislation are restricted by (1) its broad definition, (2) the First Amendment, and (3) likely applications against minority groups.

Despite the lack of criminal legislation around hate speech specifically, there are a range of other legal tools available to target similar inflammatory and dangerous speech online. Many of these laws are problematic in that they criminalize behavior with often disproportionate penalties, yet do not take a preventative or structural approach to issues of inflammatory and hateful speech.

- Cyberbullying: A number of states have addressed cyberbullying. For example, a 2017 law in Texas, which received backlash, criminalized bullying of someone under the age of 18 online or via text messages. Another bill in Nebraska would provide materials to school districts to prevent and respond to instances of cyberbullying.

- Cyber-harassment: The federal government does not criminalize cyber-harassment, although some of the behavior could be targeted under other laws such as cyberstalking. Some states target harassment online; California's penal code criminalizes the use of electronic communication equipment to repeatedly contact someone with the intent to harass or annoy.

- Cyberstalking: States generally have anti-stalking laws that can apply to the online sphere, and the federal government has a cyberstalking statute (Title 18 U.S. Code § 2261A). Stalking differs, in part, from harassment due to the repeated nature of the communications.

- Other laws that could address online hate speech include the Violence Against Women Act, hate crime statutes, and other statutes in the U.S Criminal Code. Victims of inflammatory speech can also sue under civil law, although this remains expensive and time-intensive.

Amid fears that tech companies are not effectively monitoring their platforms, recent discussion in Congress has centered on modifying Section 230 of the Communications Decency Act of 1996. This

† A research brief prepared by Adrian Shahbaz, research director for technology & democracy at Freedom House, a nonprofit, independent watchdog organization, for the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG here: https://www.ivir.nl/twg/.

provision generally shields intermediaries (such as social media companies and website owners) from legal liability for the activities of their users, although there are exceptions for criminal and state law (e.g., on harassment, stalking, protecting children), intellectual property law, and sex trafficking law. The latter is a recent development.

- In March 2018, Congress passed the Stop Enabling Sex Traffickers Act and Allow States and Victims to Fight Online Sex Trafficking Act of 2017, or SESTA/FOSTA, intended to address sex trafficking facilitated online. It represents one of the few legislative changes to intermediary liability in recent years. However, the law has had the unintended consequence of pushing companies to preemptively remove legitimate content, and sex workers and community advocates argue that it threatens their safety since targeted platforms – such as Backpage.com and sections of Craigslist – made it possible for sex workers to flee exploitive situations, communicate with one another, and build protective communities. The only two Senators to vote against SESTA were Ron Wyden (D-Ore.) and Rand Paul (R-Ky.).

- Sen. Wyden, a coauthor and staunch supporter of Section 230 as a fundamental pillar of the internet, argues that companies have embraced only the part of the law that provides protection against liability and have not actively used their censorial discretion to remove unwanted or illegal content.

- Wanting to modify the law, Sen. Mark Warner (D-Va.) published a 2018 white paper suggesting that platforms should be ordered to remove content after a court deems it defamatory or invading privacy, among other material.

- In a more aggressive approach against Section 230, Sen. Ted Cruz (R-Texas) argues that the provision requires that platforms be "neutral public forums," which could disqualify companies like Facebook from receiving special immunity if they act as "political speakers" when, as he contends, they operate with anti-conservative bias.

- Similarly, Rep. Devin Nunes (R-Calif.), who filed a lawsuit against Twitter, argues that Section 230 should not apply to the platform because it is a content creator and is politically biased against him. Likewise, Sen. Josh Hawley (R-Mo.) has also raised "viewpoint discrimination" in efforts to change Section 230.

When interpreting laws that relate to online speech, courts have generally upheld First Amendment protections. For example, in the 2014 case *Elonis v. United States*, the Supreme Court reversed the conviction of Anthony Elonis for threatening to kill his ex-wife. The conviction hinged on Elonis' threatening Facebook comments about his ex-wife, colleagues, a kindergarten class, local police, and an FBI agent. Specifically, the Court reversed the standard that allowed for criminal liability if a "reasonable person" would understand the accused's words as a threat, but ruled narrowly to only address principles of the accused's intent and not questions around whether there are "true threats" of violence.

Disinformation and foreign propaganda

A number of legislative efforts at both the federal and state levels have targeted foreign disinformation by promoting greater transparency around online advertising and foreign news, as well as by promoting digital media literacy.

Senators Amy Klobuchar (D-Minn.) and Warner, with endorsement from Sen. John McCain (R-Ariz.), introduced the Honest Ads Act in October 2017, which would require those who purchase and publish online political advertisements to disclose information about the ads to the public. In April 2018, Twitter announced its support of the act. The specifics of the bill include:

- Incorporating paid internet and digital advertisements in the definition of electioneering communication under the Bipartisan Campaign Reform Act of 2002
- Forcing platforms and websites with over 50 million unique visitors each month to publicly document people or groups spending more than $500 on political ads
- Mandating platforms to make "all reasonable efforts" to not allow foreign individuals and groups to advertise online

There have been a number of legislative efforts targeting disinformation, or viral deception, originating from foreign actors. For example, the FY 2017 National Defense Authorization Act (NDAA) incorporated the Countering Disinformation and Propaganda Act. The text created the Global Engagement Center, an interagency body housed in the Department of State that coordinates counter-propaganda efforts across the government, and also provided grant opportunities for civil society groups to work on related issues. The FY 2018 NDAA, building off its 2017 version, again included components aimed at countering foreign propaganda and disinformation. The FY 2018 omnibus appropriations bill included $250 million for a new "Countering Russian Influence and Aggression Fund." The FY 2019 omnibus increased this amount to $275 million.

At least 24 states have introduced bills establishing a council or committee focused on comprehensive media literacy education. For example, in September 2018, California passed a law encouraging media literacy in schools by forcing the state's Department of Education to provide online resources on best practices to analyze and evaluate the news. Similarly, a 2017 Connecticut law created a council in their Department of Education addressing digital citizenship, internet safety, and media literacy. In another example, a bill in Florida would require public schools to teach fifth and sixth graders how to responsibly use social media.

There have also been renewed efforts to enforce or update the 1938 Foreign Agents Registration Act (FARA) in a bid to increase transparency around the foreign funding of media outlets. Al Jazeera, RT and Sputnik, China Daily, Korean broadcaster KBS America, and Japanese broadcaster NHK Cosmomedia have registered under the law. Some civil society organizations have criticized the use of FARA against the media, noting that it could lead to politicized targeting of outlets.

Civil society initiatives

The private sector and civil society have been more actively engaged in tackling these issues. A joint Stanford-Oxford report contains a helpful primer on "What Facebook Has Done" on content moderation and News Feed controls, and WhatsApp (owned by Facebook) took measures aimed at combating the viral spread of false information through limiting users' abilities to bulk forward messages. Google also announced a series of actions to increase the integrity of news displayed on its platform. Civil society action can be categorized into new fact-checking initiatives and partnerships, increased investment and training on how reporters can verify user-generated content, and programs to increase digital media literacy among the population. There are also well-funded initiatives like the Credibility Coalition and First Draft that aim to establish standards for online content, provide educational resources, and conduct empirical research on best practices for combating misinformation.